January 2015

UNDERSTANDING STUDENT BEHAVIORS USING IMMEDIATE FEEDBACK FEATURES IN A BLENDED LEARNING ENVIRONMENT

Xin Chen
Purdue University

Graduate School Form 30
Updated 1/15/2015

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Xin Chen

Entitled

Understanding Student Behaviors Using Immediate Feedback Features In A Blended Learning Environment

For the degree of Doctor of Philosophy

Is approved by the final examining committee:

Jennifer DeBoer
Chair

Sean Brophy

Lori Breslow

Ruth Streveler

To the best of my knowledge and as understood by the student in the Thesis/Dissertation Agreement, Publication Delay, and Certification Disclaimer (Graduate School Form 32), this thesis/dissertation adheres to the provisions of Purdue University's "Policy of Integrity in Research" and the use of copyright material.

Approved by Major Professor(s): Jennifer DeBoer

Approved by: Ruth Streveler 11/13/2015

Head of the Departmental Graduate Program Date

UNDERSTANDING STUDENT BEHAVIORS USING IMMEDIATE FEEDBACK FEATURES IN A BLENDED LEARNING ENVIRONMENT

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Xin Chen

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

December, 2015

Purdue University

West Lafayette, Indiana

To my beloved Zhiwei Zhang

To my dear father Xu Chen and mother Huakai Ge

ACKNOWLEDGEMENTS

I would like to sincerely thank Dr. Jennifer DeBoer, Dr. Ruth Streveler, Dr. Sean Brophy, and Dr. David Radcliffe for recognizing my strengths and supporting me through the most dramatic shift in the course of my Ph.D. study. I would like to again thank Dr. Jennifer DeBoer for guiding me through this research. I feel extremely blessed to be able to meet you and work with you.

I would also like to thank Dr. Lori Breslow and her colleagues at the Massachusetts Institute of Technology (MIT) for providing data and information that make this research possible and for giving comments and feedback during the ongoing investigation. I also recognize the contributions of undergraduate researcher Xingyu Zhou in data cleaning and organization.

I would like to thank my dearly beloved Zhiwei Zhang for the love, companion, support, and wisdom he gave me. Thanks to my parents Xu Chen and Huakai Ge for always being there for me. I give special thanks to my best friend Hong Chen who has inspired and continues to inspire my soul.

Thanks to many others who have left either positive or negative dents in my life.

TABLE OF CONTENTS

Page

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

Appendix Figure

ABSTRACT

Chen, Xin. Ph.D., Purdue University, December 2015. Understanding Student Behaviors Using Immediate Feedback Features in a Blended Learning Environment. Major Professor: Jennifer DeBoer.

Feedback serves to close the gap between learners' current understanding and the desired understanding. Informative feedback can keep students from holding onto misconceptions, actively engage learners in knowledge acquisition, and increase confidence and motivation to learn. Yet, in the context of higher education, it is usually not possible for instructors to provide timely feedback to every individual student. This is especially difficult in first-year foundational courses due to the large number of students. Online learning platforms offer a solution by providing students immediate feedback during the course of their interactions with formative assessment tools (e.g., online homework, quizzes, embedded questions in lecture videos). For example, an automatic feedback feature of a course platform can check immediately whether a student's solution to an online homework problem is correct or not, provide an explanation of the correct answer, and point to useful resources. However, how students choose to interact with these features, and how these features influence students' learning experiences have not been well understood. Even less is known about student behaviors with these immediate feedback features in a blended learning class. Fortunately, there is now a mechanism to address these questions since the ever-increasing usage of online learning platforms such

as edX (www.edx.org), Coursera (www.coursera.org), and FutureLearn (www.futurelearn.com) enables the detailed recordings of student activities (e.g., usage logs, message streams, mobile device data). These large-scale data allow researchers to understand student behaviors in ways that were not possible before.

This study helps us begin to understand student behaviors using immediate feedback features in a blended learning environment utilizing rich quantitative and qualitative data including server logs, survey results, interviews, and video data from observations of students' problem-solving processes.

The course studied here is an introductory physics course PHYS101 (pseudonym) "Classical Mechanics" offered to all first-year students at a top tier private research university in the Northeastern United States during fall 2014. This course is a general requirement for all students in that university. This class used a blended learning format, because it had a significant online component built off of the edX platform, while students, instructors, and teaching assistants still met face-to-face regularly in lectures and problem-solving sessions. The edX course platform served as a central place hosting resources available to the class. Most importantly, the course platform offered an immediate feedback functionality called "checkable answers" as a tool in the formative assessment tasks the students performed in the course (i.e., pre-class reading questions, online homework problems, and written homework problems).

Although PHYS101 is a physics course, it is a foundational course required for all first-year students. These students could be majoring in a number of STEM fields afterwards, and PHYS101 is often an important part of their knowledge base going forwards. Therefore, PHYS101 provides a relevant context for inquiry about student

behaviors using immediate online feedback in a blended learning environment in a foundational STEM course. In this study, we address three major research questions. The first two questions focus on student behaviors while working on the online homework problems. The third question digs into the differences in student behaviors among two different types of assessment tasks (i.e., online homework problems and written homework problems):

1. What kinds of behaviors do the students demonstrate while using the "checkable answers" functionality with online homework problems?
2. What other behaviors associated with the "checkable answers" usage behaviors do students demonstrate in an online problem-solving session (e.g., accessing other online resources)? How do these other online behaviors together with the "checkable answers" usage behaviors predict student academic performance?
3. Do students demonstrate different behaviors towards the "checkable answers" functionality while solving the online and written homework problems?

Detailed tracking log data which record every time the students click on the resources on the edX course website were collected for the whole semester. 474 students were enrolled in this course in fall 2014, and the students' interactions with the course website resulted in over 30 million tracking logs during the course of a semester. This is the primary dataset in this study. A series of statistical and data mining techniques including correlation analysis, agglomerative hierarchical clustering, K-means clustering, and multiple linear regression were used to recognize the students' common behavioral patterns and understand their correlations, if any, with the students' course performance.

In addition, a survey on students' usage and perception of the "checkable answers" function was administrated in November 2014. Qualitative data including semi-structured interview data and observation videos of students' solving homework problems on the course platform were also collected. These qualitative data were used as a supplement to provide information on student background factors and deeper insights in answering each research question. For example, when answering the second research question, the problem-solving observation data enhanced the understanding of what off-site (e.g., Google, Wikipedia) and off-line (e.g., notes, paper textbook) resources students accessed as they were working on the homework problems. These findings substantively enhanced our understanding of how and why students interacted with the course platform in the ways they did in this blended learning context.

This research utilizes rich quantitative and qualitative data to address the need to understand the potential uses and effects of immediate feedback in a blended learning setting. The implications are two-fold: (1) understanding of student behaviors with immediate feedback provide instructors an anchor to give timely interventions and recommendations on students' study strategies thereby improve instructional quality, and (2) although the subject studied here is physics, it is a required course for all first-year students who are likely to be majoring in a number of STEM fields afterwards, so the results can inform the pedagogical design of other first-year STEM classes that utilize a blended learning format.

CHAPTER 1. INTRODUCTION

1.1 Statement of the Problem

Numerous factors influence student academic achievement. According to a synthesis of over 800 meta-analyses (Hattie, 2013), these factors spread across six areas: the student, the home, the school, the curricula, the teacher, and teaching and learning approaches. This synthesis identified over 100 different factors that influence student achievement, and feedback was among the top 10 most influential ones. Hattie and Timperley (2007) conceptualize feedback as "information provided by an agent (e.g., teacher, peer, book, parent, self, experience) regarding aspects of one's performance or understanding." Students often learn a concept by reconciling the new concept with their prior knowledge, or they master a skill by applying it in practice (Bransford, Brown, & Cocking, 1999). If no feedback is provided in the process, repeated erroneous practice could reinforce prior misconceptions and result in minimal learning (Pellegrino, Chudowsky, Glaser, & National Research Council, 2001). Timely and informative feedback can reduce the discrepancy between students' current understanding and the desired understanding or the desired goal (Hattie & Timperley, 2007). Feedback can also actively engage the students in information seeking and processing and enhance self-confidence and motivation to learn (Epstein et al., 2010). However, a problem arises within higher education: it is nearly impossible for instructors to provide timely feedback

to every individual student. This is especially difficult in first-year foundational and required courses, because there are often too many students in one classroom.

The proliferation of online learning environments offers a solution to the above problem. That is, the course platform can provide learners automatic immediate feedback during the course of their interactions with formative assessment tools (e.g., online homework, quizzes, embedded questions in lecture videos). For example, the course platform can check immediately whether a student's answer to an online reading question is correct or not, provide an explanation of the correct answer, and point to useful resources. However, the affordance of the designed feedback features does not always lead to the desired usage behaviors. Students can use the feedback features asynchronously in very self-directed ways, which adds a layer of uncertainty to our understanding of the impact of feedback via an online platform on learning outcomes. Literature on student behaviors as they use intelligent tutoring systems (ITS) has even documented that students may intentionally misuse the feedback and help mechanism to trick the system into giving out the correct answers (Aleven & Koedinger, 2002; Baker, Corbett, Koedinger, & Wagner, 2004). Although these studies were conducted in a different context from that of the current study, they nevertheless remind us that students in online learning environments could demonstrate various types of behaviors when using the immediate feedback features—intended or unintended by the course instructors and platform designers. What these various behavioral patterns are and how they have influenced students' learning experiences and academic performance have not been well understood.

Even less is known about student behaviors using the online immediate feedback features in a blended learning environment. A blended learning class in higher education context is one that integrates online learning experiences with face-to-face classroom learning experiences. In a blended learning class, students' experiences with the online learning component and their classroom experiences are woven together. The immediate feedback given automatically by the online formative assessment tools are likely in addition to various other forms of feedback that students receive in the classroom, in tutoring sessions, and in learning communities if they belong to one. As blended learning classes proliferate in higher education, empirical studies have reported mixed evidence of the effectiveness of blended learning formats (Garrison & Kanuka, 2004; Means, Toyama, Murphy, Bakia, & Jones, 2009; Zhao & Breslow, 2013). Bernard, Borokhovski, Schmid, Tamim, and Abrami (2014) argue that studies merely answering "big" questions such as "Is blended learning more effective than traditional classroom-based learning?" yield no actionable insight for educational practices, because these studies are often so plagued with confounds that it is hard to know which component or combination of components make online or blended learning effective or not effective. These studies, hence, cannot effectively make actionable recommendations on "do's" and "don'ts" of instructional and pedagogical design.

Therefore, there is a great pressing need to understand in-depth how each component in a blended learning class affects students' learning outcomes. Our study takes a very focused approach to address the need to understand student behaviors using one online feedback function, which primarily provides students immediate feedback on the correctness (a.k.a., corrective feedback) of their solutions to online formative

assessment tasks in a blended learning class. We analyze student behaviors using the immediate online feedback, situate these behaviors with their integrated experiences in the blended learning environment, and provide actionable recommendations on productive learning behaviors thereby to improve pedagogical design.

In the next two sections, we describe in detail the blended learning class studied here and present the specific research questions we aim to address.

1.2 PHYS101 Context

The course studied here is an introductory physics course PHYS101 (pseudonym) "Classical Mechanics" offered to all first-year students at a top tier private research university in the Northeastern United States. The Classical Mechanics course is a general requirement for all students in this university. Every first-year student has to take a Math diagnostic test. As a result, the physics department recommends some students to be placed into a more advanced version of the Classical Mechanics class, and some students to take a less difficult version of the class. Other students may decide to take the class in one of four first-year learning communities the university offers. The majority of the students enroll in the mainstream version of the class studied here – PHYS101. Most of the students taking PHYS101 are first-year students. However, there are also occasionally upper-level students taking this course either because they did not pass it the first time they took it or because they have not fulfilled this general requirement. A total of 474 students were enrolled in this course in fall 2014. These students were divided into seven sections taught by seven different instructors. Although the whole class followed the same syllabus, different instructors might have different teaching styles in terms of lecturing and arranging classroom activities.

The course relied on an active learning pedagogical model known as Studio Physics. Studio Physics loosely denotes a format instituted in 1994 at Rensselaer Polytechnic Institute by Wilson (Wilson, 1994). This pedagogy has been modified and elaborated on at a number of other universities, notably in North Carolina State University's Scale-Up program (Beichner et al., 2007; Beichner & Saul, 2003). In the classroom (studio), groups of nine students sit at a round table and collaborate in teams of three. The Studio Physics model aims to transform how introductory physics courses are taught by enabling students to be in a highly interactive, collaborative, and hands-on learning environment.

The PHYS101 class met three times a week. Two of the meetings included lectures with simple hands-on activities (e.g., clicker questions and desk-top experiments) interspersed within the lectures to help solidify the students' understanding of the concepts. The third class meeting was used entirely for team-based problem-solving activities. The third class meeting happened on every Friday, thus was referred to as "Friday problem-solving session." This Studio Physics model was piloted in another required introductory physics course, PHYS102 (pseudonym) "Electricity & Magnetism," dating back to 2001. Since then, it has been through several iterations and improvement cycles.

In addition to the active learning classroom experiences, the PHYS101 course studied here included an online component built off of the edX platform (www.edx.org). EdX is a non-profit online initiative founded by MIT and Harvard University. The edX platform has been adapted by various institutions to offer open courses to the world or as course sites for residential students. PHYS101 used the adapted edX platform as its

course platform. The course platform served as a central place for course learning resources, including homework problems, pre-class readings and questions, e-textbook, lecture slides, notes for in-class problems, notes for Friday problem-solving sessions, problem-solving videos, grade book, progress tracker, office hour calendar, and a link to the Piazza discussion forum. Most importantly, the course platform offered an immediate feedback functionality called "checkable answers" as part of the online formative assessment tools (pre-class reading questions and homework problems). The "checkable answers" provided students automatic feedback on their activities performing the formative assessment tasks outside of the classroom. When students put in an answer to a homework problem, it gave students a green check mark if the answer was right or a red "X" mark if it was wrong (see Figure 1.1).

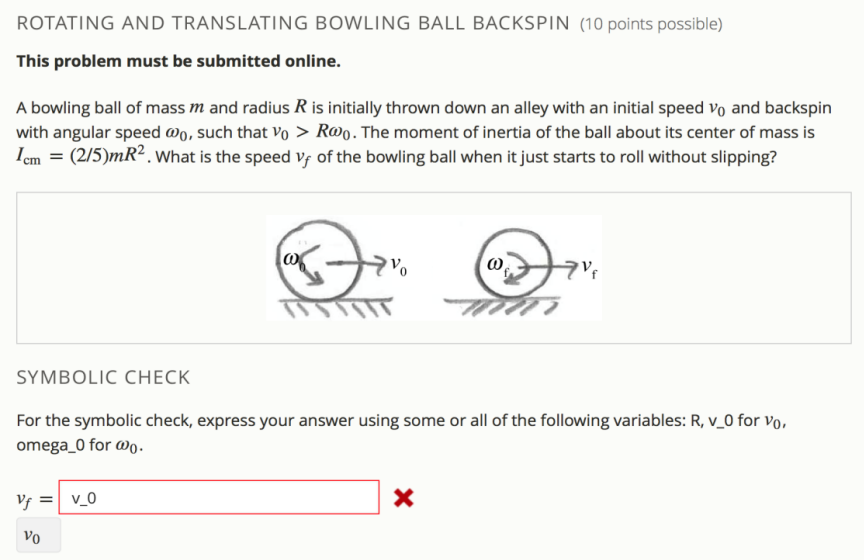


Figure 1.1 Screenshot of "checkable answers" for an online homework problem in the PHYS101 course platform

In the spring of 2014, the PHYS102 "Electricity & Magnetism" course first piloted using the edX online platform in addition to the previously established active learning classroom experiences. In a survey done near the end of that semester, 95% of the survey respondents answered "Yes" to the question of whether the online platform should continue to be used in the course. In particular, students raved about the benefits of the "checkable answers." Out of the 573 respondents, 79% of them rated the "checkable answers" on the written homework problems as "extremely useful" and 13% rated it as "very helpful" (reference blinded). These survey results serve as one motivating factor for our current in-depth study of students' experiences with the course platform in PHYS101, offered in the fall of 2014.

As described above, this class had a significant online component while students, instructors, and teaching assistants still met face-to-face regularly in lectures, problem-solving sessions, office hours, and tutoring sessions. This class thus provides a suitable context for inquiry about student behaviors using online immediate feedback features in a blended learning environment where a foundational STEM course is taught.

1.2.1 PHYS101 Assessments

PHYS101 had several formative assessments as well as summative assessments such as mid-term and final exams. In this dissertation, we intend to capture students' interactions with "checkable answers" while they were working on the two formative assessment tasks—online homework problems and hand-written homework problems. For both types of homework problems, students were allowed an unlimited number of times to check using the online "checkable answers" whether their answers were correct. However, no answer or explanation would be provided before the due date. For the online

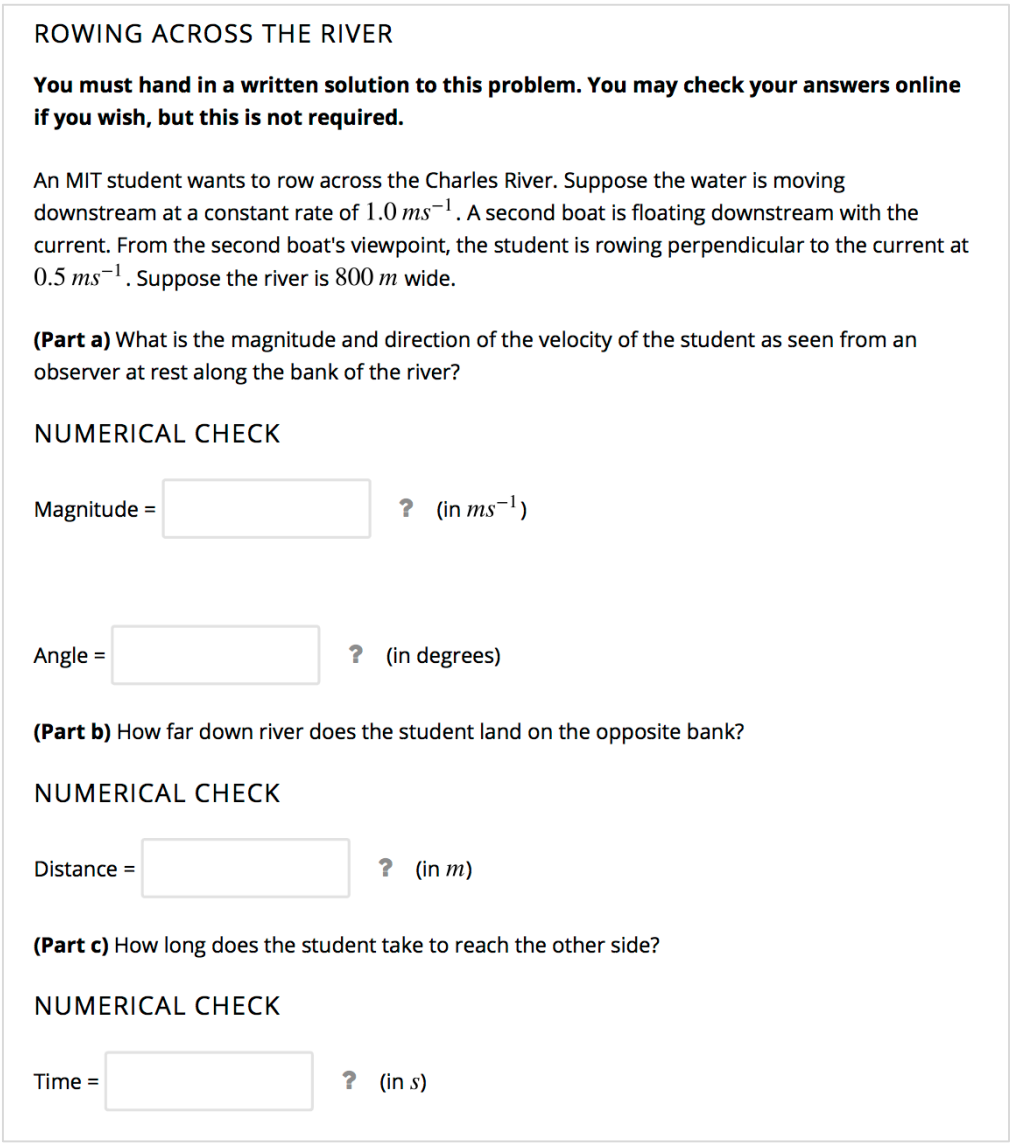
homework problems, the students only needed to input their final answers online, and they could check whether their final answers were correct or not. The hand-written homework problems were usually harder than the online homework and contained more steps. The solution for each step often required information from previous steps. Figure 1.2 is an example of a hand-written homework problem. The students were required to submit offline hand-written solutions with intermediate steps, and they were graded not only on the correctness of their solutions, but also on their hand-written problem-solving processes. Students could optionally check online to see whether their answers for each step as well as their final answers were correct. They could also choose to skip certain steps or not check their answers online at all. Again, in this dissertation, our focus is students' interactions with the online answer checker; therefore, our analyses do not involve the students' actual hand-written pieces of the homework. Instead, we use their interactions with the "checkable answers" to infer the students' attitudes, study strategies, and problem-solving processes.

The course lasted for 15 weeks, and the students were assigned 1 or 2 online homework problems and 5 or 6 written homework problems during each of the first 12 weeks. In total, there were 22 online homework problems and 63 written homework problems throughout the semester. The online homework counted for 2% of the cumulative score, and the written homework counted for 8% of the cumulative score.

Other formative assessments in PHYS101 included concept questions, experiment reports, Friday problem-solving sessions, and pre-class reading questions, which counted for 5%, 2%, 8%, and 5% respectively of the cumulative score. Concept questions were peer-instruction clicker questions, and the students were graded on participation only.

Correct answers to the pre-class reading questions were available to the students, so in this sense, the students were also only graded on participation for the pre-class reading questions.

The summative assessments included three mid-term exams and one final exam, which counted for 45% and 25% respectively towards the cumulative grade. In our data analyses, we use the final exam score and cumulative grade as the outcomes of the course.



ROWING ACROSS THE RIVER

You must hand in a written solution to this problem. You may check your answers online if you wish, but this is not required.

An MIT student wants to row across the Charles River. Suppose the water is moving downstream at a constant rate of $1.0\ ms^{-1}$. A second boat is floating downstream with the current. From the second boat's viewpoint, the student is rowing perpendicular to the current at $0.5\ ms^{-1}$. Suppose the river is $800\ m$ wide.

(Part a) What is the magnitude and direction of the velocity of the student as seen from an observer at rest along the bank of the river?

NUMERICAL CHECK

Magnitude = ? (in ms^{-1})

Angle = ? (in degrees)

(Part b) How far down river does the student land on the opposite bank?

NUMERICAL CHECK

Distance = ? (in m)

(Part c) How long does the student take to reach the other side?

NUMERICAL CHECK

Time = ? (in s)

Figure 1.2 Example of a hand-written homework problem in PHYS101

1.3 Research Questions

This study answers three major research questions. The first question focuses on understanding student behaviors using the "checkable answers" function while working on the online homework problems. The second question focuses on student behaviors other than using the "checkable answers" while working on the online homework problems. These "other behaviors" (e.g., they accessed the e-textbook after they got a red "X" mark indicating that their answer was wrong) happen in between problem checks, and thus they are proxies for students' activities before and after checking and their problem-solving strategies. We focus the first two questions on online homework problems due to (1) "checkable answers" for the online homework problems provide the most basic corrective feedback and (2) students only need to submit their answers online so that large portions of their problem-solving activities are likely to be online captured by the server tracking log data, which provide us a rich dataset to start our in-depth and focused investigation on the most basic form of automatic feedback. The third question then digs into the differences in student behaviors while working on the online versus written homework problems:

1. **What behavioral patterns do the students demonstrate while using the "checkable answers" feature with online homework problems?** This question focuses on general statistics summarizing student behaviors using the "checkable answers" while solving the online homework problems. For example, how many times do the students check before giving up or until getting the correct answers? How, if at all, are these behaviors related to the students' final exam or cumulative grades?

2. **What other behaviors associated with the "checkable answers" usage behaviors do students demonstrate in an online problem-solving session? How do these other online behaviors together with the "checkable answers" usage behaviors predict student academic performance?** To answer this question first requires identifying an online problem-solving session, and then we can say that behaviors happening within one session are associated behaviors. Using the server tracking log data, we clustered students' periods of intensive interactions with the online homework problems as problem-solving sessions. The specific clustering method is detailed in CHAPTER 3. One example question we aim to answer here is what other online resources (e.g., e-textbook, problem-solving videos) the students access when working on the online homework problems?
3. **Do students demonstrate different behaviors towards the "checkable answers" functionality while working on the online versus written homework problems?** This question examines the nuanced distinctions among the two feedback settings in terms of the students' perception of, attitudes towards, and behaviors with the feedback. For example, do students demonstrate distinct or similar behaviors with the online and written homework problems? If so, what are the possible underlying reasons?

To answer these research questions, we collected various types of data in order to understand the students' behavioral patterns using the immediate feedback features and how these behaviors are related to their learning outcomes. First and foremost, the edX platform has the capacity to record every interaction the students had with the course site.

These rich server tracking log data allow us to identify students' behavioral patterns using large-scale data mining techniques. The server tracking log data for the 474 students constitute the primary dataset for this dissertation. As supplements, we also collected surveys, interviews, and observation videos of students solving homework problems on the edX platform. These additional data help us better understand students' perceptions, attitudes, and self-efficacy factors related to using "checkable answers."

1.4 Overview

The next chapter discusses the literature within which we situate this study. Relevant literature includes empirical studies of blended learning environments, theory and empirical research on the impact of feedback on learning, and student behaviors and study strategies using computer-mediated feedback. In particular, we describe a conceptual model of the impact of feedback on student achievement. Chapter 3 then elaborates on how we operationalize the key constructs of the feedback model in the PHYS101 course context, and the sampling, data collection, and data analysis methods. The methodological framework involves a mix of inductive and deductive quantitative methods. Chapter 4 describes the results obtained from analyzing the tracking logs and the "checkable answers" survey responses. Chapter 5 discusses the implications of the results, and Chapter 6 discusses the limitations of this research. Finally, Chapter 7 concludes this dissertation with potential future work.

CHAPTER 2. RELEVANT LITERATURE AND CONCEPTUAL MODEL

2.1 Blended Learning

Internet-based technologies are pervading higher education. Institutions and faculty across diverse disciplines are incorporating online learning into traditional lecture-based classes. Blended learning, also called hybrid learning, usually refers to the integration of face-to-face learning experiences and online learning experiences (Garrison & Kanuka, 2004; Means, Bakia, & Murphy, 2014, Chapter 3; Zhao & Breslow, 2013). While the face-to-face classroom experiences are largely synchronous, the online learning experiences outside of the classroom are mostly asynchronous and students can work with many of the materials online in a self-paced manner.

Following the method to categorize online learning proposed by Twigg (2003), Zhao and Breslow (2013) categorize the broad spectrum of blended learning formats into four major models: the replacement model, the supplement model, the emporium model, and the buffet model. The replacement model reduces lecture time by partially or fully substituting online components or interactive activities for the lecture time. The supplemental model usually retains the same amount of lecture time while supplementing the lecture with an online component where students could engage with course-related materials outside of the classroom. The emporium model relies solely on online learning with help from on-demand instructors and TAs. The buffet model provides students with

a “menu” with a mix of online and face-to-face learning activities and students can pick a combination of materials that suits their learning objectives. For example, “flipped classroom” falls into the replacement model category, where students usually learn the course materials online prior to class, and then the class time is used entirely for active learning activities. The course studied in this dissertation, PHYS101, falls primarily into the supplement model category. As described in section 1.2, the Studio Physics model used by PHYS101 has been established for more than 10 years, and the edX course platform is an online component added to this basic class structure. However, the PHYS101 blended learning environment also had a flavor of the replacement model, because active learning activities and teamwork were spread across lectures, traditional recitations were replaced by hands-on problem-solving sessions, and online homework partially substituted for paper-based homework. These four models are far from exhaustive in practice, and many blended learning formats fall somewhere in between two models or do not fit at all.

Blended learning has several potentially beneficial characteristics that are not available in traditional lecture-based classes. One example of such potential benefits is that faculty members could have access to analytics from students’ online learning activities, and thus could adjust the in-class activities based on students’ mastery level of the concepts and skills (Means et al., 2014, Chapter 3). Another benefit is that blended learning usually allows for more flexible modes of participation. In some blended learning classes such as the “flipped classroom,” students could learn the course materials outside of the classroom at their own pace, thus allowing for reduced lecture time in class. Faculty then could use the saved class time to introduce more active and engaging

learning experiences (Garrison & Kanuka, 2004). Means et al. (2014, Chapter 7) suggest that the self-paced online component combined with help from the instructor and peers in class could be particularly helpful for students with weak academic backgrounds and in remedial classes.

Because of the benefits described above, Garrison and Kanuka (2004) argue that blended learning has the potential to transform higher education by enhancing both the effectiveness and efficiency of teaching and learning, and therefore supporting deep and meaningful learning. Bele and Rugelj (2007) argue that blended learning combines the elements of online learning and classroom instruction, so it provides the potential to take "the best of both worlds."

However, this potential has yet to be fully realized. Some have found that blended learning may increase the cognitive load demanded of students (Bower, Dalgarno, Kennedy, Lee, & Kenney, 2015). Also, some instructors might be concerned that unproctored exams or homework assignments instrumented through the online platform may be biased or otherwise invalid (Ardid, Gómez-Tejedor, Meseguer-Dueñas, Riera, & Vidaurre, 2015). Empirical studies report mixed evidence of the effectiveness of blended learning formats (Bowen, Chingos, Lack, & Nygren, 2014; Lack, 2013; Zhao & Breslow, 2013). We argue that this situation is largely due to the fact that there are virtually countless instructional and platform design possibilities for blended learning classes in various contexts; the range of blended learning structures and features in practice overwhelms our ability to evaluate and make meaningful generalizations about the format. For example, a whole range of analytics on students' out-of-class activities could be provided to the faculty members, from aggregate grades to individual-level behaviors.

However, which aspects of the analytics are most needed and whether faculty members appropriately integrate the analytical results into their in-class instructional design are unclear. Often times, the analytics fail to provide more insightful value than the "sense" of how students are doing the faculty could get from in-person interactivities in the classroom, homework scores, and exams. There are also myriad possible ways that each feature in the online learning platforms could be designed and used. These factors pose a huge challenge to determining best practices for the pedagogical and platform design of blended learning environments. Studies on the effectiveness of blended learning are often convoluted by many confounding factors, and we lack a clear understanding of which particular components are functioning in what ways. As Bernard et al. (2014) puts it, and we concur, studies that answer the "big" questions (e.g., Is blended learning more effective than traditional classroom instruction?) "generally fail to establish an alignment of evidence that addresses the 'do's' and 'don'ts' of instruction via rigorous research." (p. 89)

This current study does not intend to justify whether blended learning is effective or not. Rather, a blended learning environment is the background and context underlying this dissertation. It is a relevant context to study automatic immediate feedback features for teaching and learning, which are becoming more and more prevalent across online learning platforms but are not well understood yet. Students' behaviors using the immediate feedback features online could be influenced by many aspects of their intertwined offline and online experiences in blended learning environments. We take a very focused approach to understanding and improving blended learning platforms, starting by studying the specific component of immediate feedback within a blended

learning context. In doing so, our goal is to bring rich insights to the larger scope of pedagogical design in a blended learning environment.

2.2 A Conceptual Model of Feedback

Feedback is defined as "information provided by an agent (e.g., teacher, peer, book, parent, self, experience) regarding aspects of one's performance or understanding" (Hattie & Timperley, 2007). We know from constructivism learning theory and numerous empirical studies in cognitive science that students learn a new concept or skill drawing on their prior knowledge (Black & Wiliam, 1998; Bransford et al., 1999; Sebatane, 1998). Feedback plays an important role in closing the gap between current understanding built from prior knowledge and the desired goal thereby enhancing student performance and achievement, as we illustrate in Figure 2.1. If there is a lack of prior understanding, feedback is usually ineffective, since the students do not know how to relate the new information with what is already known (Kulhavy, 1977; Kulhavy & Stock, 1989). Therefore, feedback is most powerful when it corrects misinterpretations built from prior knowledge in order to form new understanding. If a faulty interpretation is established and students keep practicing it without any feedback, this will have a small or even negative effect on learning (Pellegrino et al., 2001, Chapter 3).

Feedback is usually more effective if it focuses on the mastery of clear and specific goals. If the learning goal is not well defined or too general, feedback may unintentionally lead to reduced learning. When students are unclear about the goal or specific steps needed to reach the goal, they are likely to pursue alternative goals with much lower standards or abandon the goal completely, which makes the students

disengage from deep and meaningful learning (Hattie & Timperley, 2007; Locke & Latham, 1990, 2002).

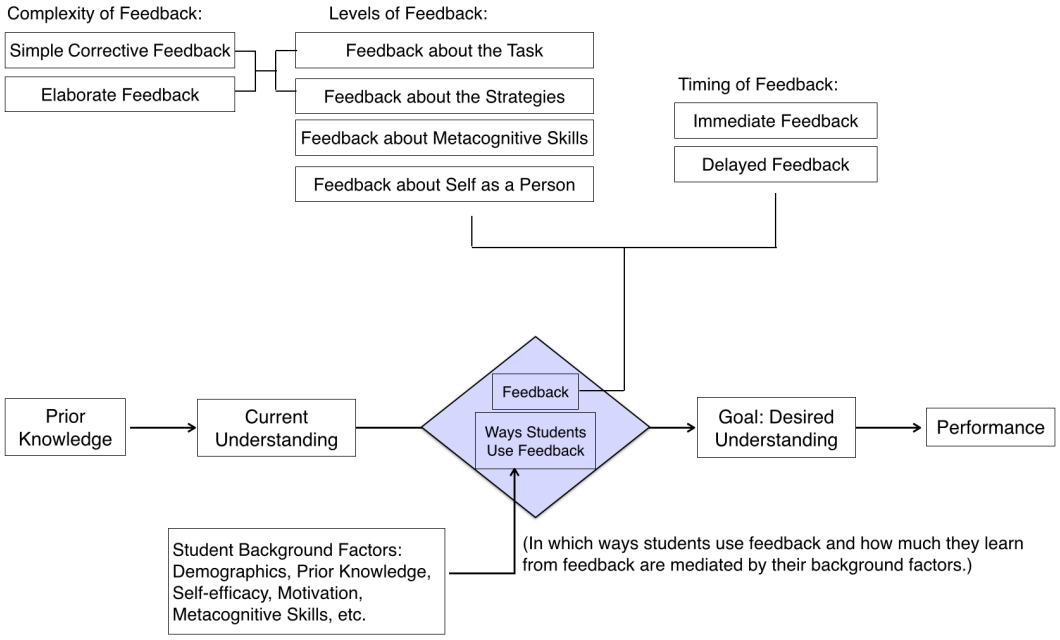


Figure 2.1 A conceptual model of feedback

According to Hattie and Timperley (2007), there are four major levels of feedback: feedback about the task, feedback about the processing of the task, feedback about self-regulation, and feedback about the self as a person. To more succinctly and precisely refer to these levels in this dissertation, we renamed "feedback about the processing of the task" and "feedback about self-regulation" to "feedback about strategies" and "feedback about metacognitive skills" respectively.

There also exists much discussion about the relative complexity (i.e., simple vs. elaborate feedback) and the timing of feedback (i.e., immediate vs. delayed feedback) and how complexity and timing affect the effectiveness of feedback (A. Butler, Karpicke, &

Roediger III, 2007; Clariana, 1999; Clariana, Wagner, & Murphy, 2000; Elder & Brooks, 2008; Shute, 2008). We elaborate further on the levels, complexity, and timing of feedback in the next three sections.

2.2.1 Levels of Feedback

The four levels of feedback are feedback about the task, feedback about the strategies, feedback about metacognitive skills, and feedback about the self as a person. As we show in Figure 2.2, these levels of feedback range from very specific information on the correctness of the task to very generic comments on the student as a person.

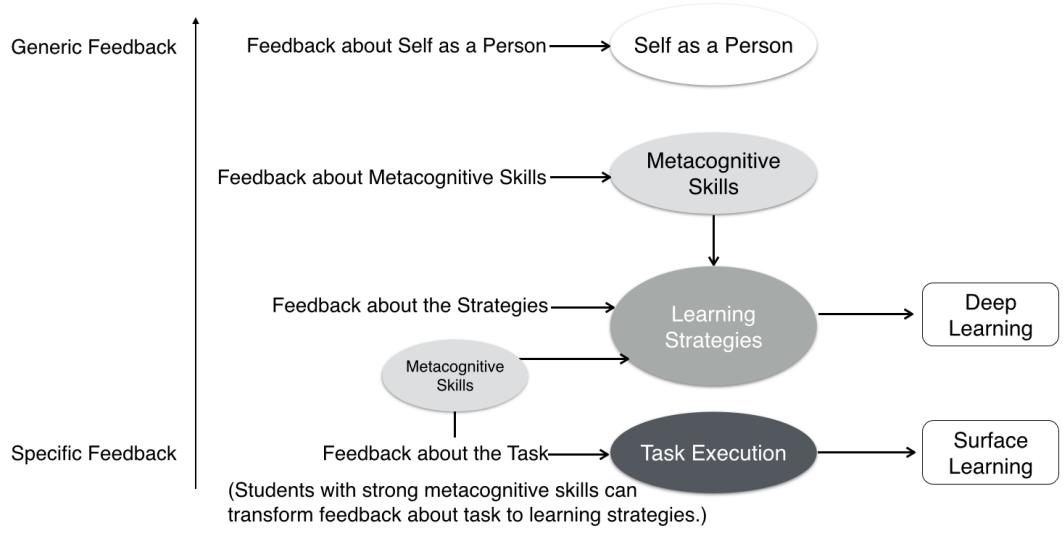


Figure 2.2 Four levels of feedback

2.2.1.1 Feedback about the Task

Feedback about the task usually refers to information about the correctness of the work (a.k.a., corrective feedback), explanations of the correct solutions, and directions to acquire more or different information in order to accomplish the task. Corrective feedback includes feedback on the correctness of each step during the execution of the

task. Feedback at the task level is usually very specific and is therefore easy to comprehend by the students. However, one drawback is that feedback about the task usually does not generalize to other tasks (Thompson, 1998), and too much feedback only at the task level can direct students' attention to trial-and-error strategies that support the realization of immediate goals rather than long-term goals (Kluger & DeNisi, 1996).

2.2.1.2 Feedback about the Strategies

Feedback about the strategies usually focuses on the underlying processes of the task or broader learning strategies. For example, "You need to draw a free-body diagram in order to understand different forces, and apply the conservation of momentum to solve this problem," or "Reviewing the reading summaries before class will help you better understand these concepts." Researchers argue that feedback about the strategies is more effective to enhance deeper learning compared with the kind of surface knowledge gained from corrective feedback (Balzer, Doherty, & O'Connor, 1989; Earley, Northcraft, Lee, & Lituchy, 1990). However, this type of feedback imposes a heavier cognitive load on students than feedback about the task. The same learning strategies might work differently for different students. Student background factors such as academic preparation, prior experiences, and individual characteristics influence what learning strategies they use and how much they learn from feedback about strategies. In particular, for students who are less experienced with the subject matter and learning strategies, appropriate learning habits and strategies take a much longer time to develop.

Instead of seeking direct feedback about the strategies, some students can transform feedback at the task level into effective learning strategies, as we show in

Figure 2.2. This is called the interactive effect between feedback about the task and the strategies (Earley et al., 1990). Feedback at the task level can help students who already demonstrate metacognitive awareness develop appropriate learning strategies on their own, which can be more powerful than direct feedback about the strategies. For example, corrective feedback can help students reject erroneous hypotheses and provide cues to identify the correct solutions. If the students are self-conscious and strategic, they can gradually develop error detection skills, which are very valuable self-feedback strategies (Harackiewicz, Manderlink, & Sansone, 1984).

2.2.1.3 Feedback about the Metacognitive Skills

Vockell (2004, Chapter 7) defines metacognition as "learners' automatic awareness of their own knowledge and their ability to understand, control, and manipulate their own cognitive processes." Vockell (2004, Chapter 7) points out that metacognitive skills also include important affective and personality components such as attitudes, commitment, confidence, and motivation. For example, a student is aware of the importance and value of mastering a concept, thus is committed to master it, and deliberately develops a positive attitude and the necessary strategies to master it. Feedback about metacognitive skills thereby aims to enhance students' self-monitoring, self-evaluation, and self-reflection abilities, as well as their commitment, confidence, and motivation to learn. For example, "You already know that you have confused the conservation of energy with the conservation of momentum in another problem before, check to see if you are making the same mistake here." We can see that this example contains some information about the strategy (e.g., use conservation of energy or

conservation of momentum), but its major focus is to direct students' attention to self-monitoring and self-evaluation. Another example is provided in (Roll, Aleven, McLaren, & Koedinger, 2011), where the authors designed an intelligent tutor agent as a plugin for any intelligent tutoring systems (ITS) to provide students immediate metacognitive feedback on their help-seeking behaviors. They found that the help tutor improved students help-seeking and self-assessment skills, and these skills were transferred to learning new domain-level content the month following the intervention when the help tutor was no longer in place.

As illustrated in Figure 2.2, metacognitive skills govern the process of transforming knowledge on task into generalizable learning strategies and habits, which further leads to deep learning. Researchers have demonstrated that metacognitive skills can be taught and learned (Kuhn, 2000; Pellegrino et al., 2001). Therefore, metacognition is also influenced by student background factors like prior exposure to training experiences on metacognitive skills. Students with less effective metacognitive skills are less competent at establishing effective learning strategies. Many researchers thus argue that the ultimate goal of teaching should not be to make students master sets of concepts and skills accurately, but rather to foster strong metacognitive skills so that students can self-monitor their problem-solving processes and self-assess their learning strategies and error-correction methods (Mathan & Koedinger, 2002; Pellegrino et al., 2001).

2.2.1.4 Feedback about the Self as a Person

Feedback about the self as a person, for example, "You are a great student," is usually very general and tangential to learning. If this praise is reworded, as "You are

great because you used this concept and applied it in an appropriate way," it could be more relevant and helpful. This example demonstrates that if feedback about the self is combined with feedback about the task, strategies, and metacognition and aims to direct attention away from the self to the task, it could be helpful for learning. However, most of the time in practice, feedback about the self is given in a form similar to the example ("You are a great student.") Researchers argue that this type of feedback is uninformative and yields no value for learning (R. Butler, 1988; Hattie & Timperley, 2007).

2.2.2 Complexity of Feedback: Simple Feedback vs. Elaborate Feedback

Another way of classifying feedback is based on its complexity. Feedback is categorized into simple feedback and complex elaborate feedback in this way (Elder & Brooks, 2008). Most of the time in practice, simple feedback only refers to information indicating whether the responses or results are correct or not (corrective feedback), while elaborate feedback can contain various forms of additional information such as explanations of why a particular answer is incorrect, what would be the appropriate next step to take, hints about useful resources, and hints about procedural skills and problem-solving strategies. Kulhavy and Stock (1989) propose to use the term "elaborated feedback" for all types and all levels of feedback that provide more information than merely the correctness of the solutions. However, in most empirical studies, elaborate feedback is often at the task level or strategy level and rarely at the metacognitive level or self as a person level (Chase & Houmanfar, 2009; Kulhavy, White, Topp, Chan, & Adams, 1985; Phye & Bender, 1989). Therefore, in this dissertation, we use "elaborate feedback" to refer to task level and strategy level feedback that consists of more information than basic corrective feedback.

Similarly as the distinction between task level feedback and strategy level feedback mentioned in section 2.2.1.2, elaborate feedback is also usually taxing on working memory and can impose heavier cognitive load to the students than simple corrective feedback. Hence, it raises a potential risk of deviating students' attention and interfering the problem-solving process. Providing the elaborate feedback information stepwise in manageable pieces and offering the opportunity to apply the information provided on a previous step to the next step could be an effective way to mitigate this issue (Narciss & Huth, 2004). In the written homework problems of PHYS101, one problem was segmented into several steps, and some steps required information from previous steps. Though the "checkable answers" only provided simple corrective feedback for each step, it is still more elaborated than providing simple corrective feedback to just one final answer.

Not all students can benefit from elaborate feedback due to variations in their academic preparation and other background factors. For example, Gordijn and Nijhof (2002) conducted a study in a technology education course and found that students with strong reading compression skills benefit from elaborate feedback significantly more than students with poor reading skills.

2.2.3 Timing of Feedback: Immediate vs. Delayed Feedback

Immediate feedback is provided right after responses are given by the students, while delayed feedback could occur minutes, hours, days, or even longer after the initial responses. Much debate exists in the literature over the effectiveness of immediate versus delayed feedback. Proponents of delayed feedback refer to the superiority of delayed over immediate feedback as the Delay-Retention Effect (DRE). DRE was observed as early as

the 1960s (Brackbill, Bravos, & Starr, 1962; Brackbill & Kappy, 1962). Kulhavy and Anderson (1972) propose an interference-preservation hypothesis to explain DRE, stating that if the feedback were delayed, the initial incorrect response would likely have been forgotten during the delay interval and thus would not interfere with the delayed feedback helping to establish correct understanding.

However, many recent studies argue against DRE. Kulik and Kulik (1988) report that delayed feedback appears to be more effective in experimental studies in the laboratory, but applied studies in the classroom usually produce the opposite results. Other studies demonstrate that immediate feedback reduces preservation of initial incorrect responses and also increases students' confidence and motivation to learn (Dihoff, Brosvic, & Epstein, 2003; Dihoff, Brosvic, Epstein, & Cook, 2004). These authors argue against DRE saying that the benefits of DRE are likely related to the general benefits of feedback rather than the delayed timing. Immediate feedback is sometimes coupled with an answer-until-correct (AUC) procedure. In an AUC task, students are given feedback immediately on the correctness of their answers and are required to work on the problem until they get the correct answer. It is found that when coupled with the AUC procedure, immediate feedback is even more effective, because it actively involves students in information processing while seeking the correct answer (Brackbill, Adams, & Reaney, 1967; Epstein & Brosvic, 2002; Wilcox, 1982).

In an effort to reconcile the inconsistent results about the effectiveness of immediate versus delayed feedback, researchers have proposed a variety of hypotheses. For example, Shute (2008) synthesizes findings from Corbett and Anderson (2001) and Schroth (1992) and proposes that delayed feedback maybe more effective to promote

transfer of learning especially on concept-formation tasks, while immediate feedback is more effective on short-term procedural tasks. This is consistent with another proposition that immediate feedback may be more effective at the task level, while delayed feedback could be more powerful at the learning strategy level (Clariana et al., 2000; Hattie & Timperley, 2007; Schmidt, Young, Swinnen, & Shapiro, 1989). Another such effort to reconcile competing hypotheses about feedback timing is documented in Mathan and Koedinger (2002). After reviewing various studies on the timing of feedback, the authors conclude that the effectiveness of feedback depends less on timing than on the nature of the task and the metacognitive capability of the individual student. This conclusion is consistent with what we have mentioned that in what ways the students use the feedback, how much they could learn from the feedback, and whether they could establish appropriate learning strategies from the feedback are influenced by background factors such as prior knowledge, demographics, self-efficacy, motivation, metacognitive skills, etc. We briefly describe these factors in the next section.

2.3 Student Background Factors

Student achievement is influenced directly or indirectly by numerous factors. In a synthesis of over 800 meta-analyses, Hattie (2013) identifies more than 100 factors relating to student achievement. These factors include many student background factors such as socio-economic status, gender, prior academic achievement, learning strategies, metacognitive skills, self-efficacy, motivation, personality, and personal health. Many of these student background factors are not mutually independent. They often have an effect on each other or covary with each other. For example, Coutinho (2008) reports that the relationship between metacognition and student performance is fully mediated by self-

efficacy, which means students who have strong metacognitive skills also have strong belief in their capabilities to successfully solve a problem. In addition, many of these factors are often influenced by the broad learning environment. For example, it is documented that gender stereotypes in STEM fields have formed a "chilly climate" for women and thereby negatively influence women's self-efficacy, self-confidence, and sense of belonging in these fields, which in turn negatively influence female students' achievement in STEM fields (Hall & Sandler, 1982; Riley & Pawley, 2011; Zeldin & Pajares, 2000).

A growing body of literature shows evidence that the ways in which students interact with technology, and, more specifically to this dissertation, the ways in which students use automatic feedback provided by technology and how much they learn from the feedback, are also influenced by background factors such as academic preparation, demographics, socio-economic status, study environment, computer proficiency, metacognitive skills, commitment, confidence, motivation, and emotional experience (Halder, Saha, & Das, 2015; Mathan & Koedinger, 2002). For example, Mathan and Koedinger (2002) show that students with high computer proficiency achieve an equal level of performance regardless of whether they are provided with immediate feedback on task correctness or delayed feedback on self-monitoring and error-detection skills, while students with low computer experience perform significantly better when delayed feedback is provided. Timmers, Walraven, & Veldkamp (2015) find in their study that given computer-based regulation feedback, performance improves only for female students.

According to social constructivism learning theory (Kukla, 2000; Vygotsky, 1978, 1986), learning is a complex process that happens in a social context, and knowledge is socially constructed. Not only do an individual student's background factors influence learning outcomes, but other socio-environmental factors such as peer influence, collaborative learning strategies, and teacher-student relationships also have an effect on learning.

In this dissertation, we consider many of these background and environmental factors as important control variables or mediator variables. We will detail how we operationalize these factors and other key concepts in section 3.1 of the METHODS Chapter.

2.4 Diverse Student Behaviors with Computer-Mediated Feedback

In previous sections, we have discussed extensively the theoretical and empirical studies about the effect of feedback on learning. We have come to the understanding that what types of feedback are provided to the students, how much prior domain knowledge the students have, how much cognitive and metacognitive capability the students possess in order to transform the feedback into effective learning, and various other background factors all have an influence on student achievement. Even if we provide the right level of feedback at the right time, there can still be numerous ways in which students interact with these feedback features online outside of the controlled classroom environment. Students can use the feedback features in very autonomous and asynchronous ways, which adds a layer of uncertainty to our understanding of the impact of feedback via an online platform on learning outcomes. The affordance of the designed feedback features does not always lead to the desired usage behaviors.

Several studies have researched students' various behaviors using immediate feedback and help features in ITSs (Baker, Corbett, Koedinger, et al., 2004; Roll, Baker, Aleven, & Koedinger, 2014). The following are three broad categories of behaviors identified when students use feedback features in ITSs, where the first two are usually referred to as "gaming the system":

1. Hasty trial-and-error behaviors, avoiding or ignoring help
2. Help abuse, tricking the system to give out the correct answers
3. Desired feedback and help seeking behaviors

Baker, Corbett, Koedinger, et al. (2004) found that, controlling for prior knowledge, gaming the system was strongly correlated with lower post-test scores, and students who gamed the system were more likely to do so on more challenging problems. In a subsequent study (Baker, Corbett, & Koedinger, 2004), utilizing server log data that track students' actions in the ITS, Baker and colleagues built a computational model to automatically detect students who gamed the system and were hurt by this behavior.

Research on students' help seeking behaviors found that seeking the right level of help at the right time improves learning (Newman, 1994; Ryan, Patrick, & Shim, 2005). Roll et al. (2014) found that help-seeking behavior on difficult problem steps was correlated with improved learning, and help abuse was correlated with poor learning in the ITS. Yet students with weak metacognitive capability may fail to recognize what the right level of help is for themselves (Kruger & Dunning, 1999). In addition, Roll et al. (2014) also found that for students with low prior knowledge, it was more beneficial for them to avoid help. They explained this counter-intuitive phenomenon by hypothesizing

that the students with lower prior knowledge might benefit from trial-and-error before they could make sense of the help hints.

This current dissertation builds on a relevant part of the literature about ITSs and taps into the rarely understood area of student behaviors using immediate feedback features via an online course platform in a blended learning environment.

2.5 Study Strategies

Although ITS and online learning platform are usually used in different contexts, literature on student behaviors using ITS nevertheless reminds us that students in online learning environments could also demonstrate different patterns of usage behaviors that may indicate overall productive or counterproductive study strategies.

Student study skills and study strategies have been studied by many researchers (Blumner & Richards, 1997; Richards, 2001; Weinstein, 1996; Weinstein & Hume, 1998). For example, Streveler, Hoeglund, & Stein (2003) found, from analyzing a 42-item survey on 285 engineering students' study strategies, that active learning strategy was significantly positively correlated to grade point, while test anxiety, procrastination, and lack of focus were significantly negatively correlated to performance. Ericsson revealed the important role played by deliberate practice in the acquisition and maintenance of expertise (Ericsson, 2004, 2015; Ericsson, Krampe, & Tesch-Römer, 1993).

However, measures of students' productive versus counter-productive study strategies based on their behaviors with interactive features in online and blended learning platforms are rarely discussed. One notable exception is a work-in-progress study put forth by Krumm and colleagues (Krumm et al., 2015). They seek to develop measures for students' "productive persistence"—strategic behaviors that benefit learning

in an online learning platform. The term "productive persistence" was first coined by Treisman to broadly describe the package of skills and tenacity that students need to succeed (Silva & White, 2013).

Similarly, in our study, we investigate whether we could identify students' quick and instrumental vs. strategic and productive behaviors using the "checkable answers" feature. For example, our second research question focuses on understanding students' other behaviors associated with the "checkable answers" usage behaviors. These associated behaviors include behaviors such as students accessed the online textbook after getting a red "X" indicating their answer was incorrect, which could be identified as a type of help-seeking behavior. If the students took certain steps to diagnose their errors, and eventually got to the correct solutions, we could consider these behaviors as highly strategic problem-solving approaches. We estimate the relationship between these behaviors, course performance, and self-efficacy and attitudinal measures.

In the next chapter, we delineate how we operationalize the key concepts presented in this current chapter, and give an overview of our data collection and analysis methods.

CHAPTER 3. METHODS

3.1 Operationalization of Key Concepts

Feedback must be addressed in a learning context, and, in this dissertation, our context is the PHYS101 course. In this section, we elaborate on the operationalization of key concepts of interest in the PHYS101 blended learning environment. Key concepts, as described in the above chapter, include the level, complexity, and timing of feedback, student performance, student prior knowledge, and other background factors.

In Table 3.1, the self-efficacy measure was adapted to this specific discipline from the motivation scales in the Motivated Strategies for Learning Questionnaire (MSLQ) (Pintrich, Smith, Garcia, & Mckeachie, 1993). One important background factor is participation in the pre-college bridge program, which is designed for students who feel they need additional preparation for college. This program is offered by the university's Office of Minority Education, and historically, had a large proportion of underrepresented minority (URM) participants. In this program, which is held during the summer prior to the freshmen year, students learn basic concepts and skills in Math, Physics, Chemistry, and Communication/Writing. They also attend workshops on cultivating effective learning strategies and time management skills. These students might have initially felt less prepared, but this program helps them learn crucial concepts related to early university courses and study strategies. In addition, they get familiar with

the campus environment and make a community of friends before other students arrive. As a result, students who participated in the bridge program may be equally or even more prepared than students who did not. This is one example of a key background factor and the reason why it is important to control for it.

Table 3.1 Operationalization of key concepts in PHYS101 context

Key Concept	Operationalization in PHYS101 Context
Level, Complexity, and Timing of Feedback	• "Checkable Answers" providing correctness of final answers to online homework problems (task-level simple corrective immediate feedback) • "Checkable Answers" providing correctness of each intermediate step and final answer to written homework problems (task-level stepwise corrective immediate feedback)
Student Performance	• Online and Written Homework Scores in PHYS101 • Final Exam Scores in PHYS101 • Cumulative Grades in PHYS101
Student Attitudes and Perceptions	• Self-efficacy to perform in PHYS101 (Pintrich et al., 1993) • Perception of the utility of PHYS101 for the Student's Future (Husman, Derryberry, Crowson, & Lomax, 2004) • Perception of the utility of the "Checkable Answers"
Prior Math & Science Ability	• Math Courses Taken Concurrently • SAT II Subject Test—Science
Other Background Factors	• Gender • Ethnicity • First Generation College Students or Not • Financial Aid Based on Family Income • Participation in the Pre-College Bridge Program

There are also many other background factors we could only observe from qualitative data such as students' collaborative styles when working in teams, metacognitive awareness, and student-instructor relationship. Still, some other factors are not represented in any of our datasets such as students' computer proficiency, reading

comprehension skills, and instructors' teaching styles. We discuss the potential bias posed by unobservable factors in our limitations section in CHAPTER 6.

Students' performance is operationalized as their online and written homework scores, final exam scores, and cumulative grades in this course.

In the next two paragraphs, we elaborate on the operationalization of the levels, complexity, and timing of feedback. Moreover, we have developed two broad sets of hypotheses in terms of the level and timing of feedback. These hypotheses are relevant to all of our three research questions, and, if verified, have implications for the improvement of pedagogical design.

3.1.1 Two Hypotheses

The "checkable answers" in PHYS101 provided immediate corrective feedback at the task level, that is, the platform told the students whether their answer was correct (green check mark) or incorrect (red "X" mark). Furthermore, feedback for written homework problems was stepwise. This feedback was more elaborate than simple corrective feedback, but it was still only at the task level. Notwithstanding the fact that "checkable answers" only provided task level feedback, there exists an interactive effect between feedback about the task and the learning strategies as shown in Figure 2.2. This transformative process is governed by students' metacognitive skills and is also directly or indirectly influenced by many other background factors. We therefore arrive at the first hypothesis: *we will observe diverse behavioral patterns emerging from our data mediated by diverse student background factors.* For example, some students may demonstrate highly strategic error detection behaviors—when they find out their answer is incorrect, they follow very strategic steps to diagnose their answer such as first

checking algebraic errors, then checking conceptual errors by browsing the textbook and course slides, and finally coming up with a carefully worked out new answer. Yet other students may demonstrate quicker, more instrumental guessing behaviors without carefully diagnosing the error in their answer. They may frequently guess at the right answer and use the "checkable answers" to see if their guesses are correct. These are merely two examples of possible behavioral patterns we will observe from our data, and it is likely that several other behavioral patterns will emerge and some students will demonstrate a mix of several behavioral patterns in different problem settings. These different behavioral patterns can help us understand students' overall productive or counterproductive study strategies. This understanding can provide instructors an anchor to give timely interventions and recommendations for students' study strategies, thereby improving instructional quality.

In terms of the timing of feedback, "checkable answers" provided the PHYS101 students automatic immediate feedback. For that practical reason, we focus in this dissertation on student behaviors using immediate feedback. We mentioned in section 2.2.3 that when immediate feedback is coupled with an answer-until-correct (AUC) procedure, it is usually more effective. The "checkable answers" did not force students to answer until correct, yet many students still did strive to work until they got the problem correct. We therefore arrive at the second hypotheses: *students who always try to get the correct answers perform better than students who do not*. The verification of this hypothesis can potentially provide us empirical evidence on whether to recommend that the design of the feedback functionality should include cues or messages that encourage students to work until they get the correct solutions. However, this hypothesis is complex

and challenging to immediately test, since there is likely a selection bias at play, since students choose whether or not to answer until correct. Therefore, it is important that we tease apart the relationship between students' demonstration of this behavior and their success in order to recommend constraining this feedback functionality or even directing students towards this learning strategy.

In sum, based on existing theoretical and empirical literature on the types and timing of feedback, we have developed the following two broad sets of hypotheses that are relevant to all three of our research questions:

1. **We will observe diverse behavioral patterns mediated by diverse student background factors. For some students, the corrective feedback from "checkable answers" could foster their development of productive learning strategies during the process of identifying the correct solutions. On the contrary, with solely corrective feedback, some students may demonstrate hasty guessing behaviors or counterproductive disengaging behaviors. Still other students could demonstrate a mix of strategic and instrumental behaviors.** The null hypothesis is that we may not be able to distinguish any behavioral patterns that are meaningful for educational practices, or we may not be able to identify relations between the behavioral patterns with student background factors or students' performance. This hypothesis is relevant to all of our three research questions because we expect to identify various behavioral patterns when answering all of the three research questions: "checkable answers" usage behaviors when working on the online homework problems in the first research question, other behaviors associated with the "checkable answers" usage

behaviors when working on the online homework problems in the second research question, and "checkable answers" usage behaviors with the two different types of online formative assessment tasks (i.e., online homework problems and written homework problems) in the third research question.

2. **Although not required to do so in PHYS101, students who always demonstrate answer-until-correct (AUC) behaviors on homework problems perform significantly better compared with students who do not.** The null hypothesis is that we will not find any significant difference in performance between students who always work until they get the problem correct and who do not. This hypothesis is relevant especially to the first and third research questions. We test this hypothesis with "checkable answers" usage behaviors in online homework problems in the first research question, and verify whether the same conclusion applies to written homework problems in the third research question.

The study in this dissertation uses a mix of inductive and deductive approaches. Besides testing the above two broad sets of hypothetical behavioral patterns, we may find many other patterns emerging from the data when answering each of our three research questions. We look into these categories of behavioral patterns and their relationship to students' performance.

Besides immediate feedback, the course platform also provided students delayed feedback, as solutions to homework problems were released after the due date (about 10 days after each homework was released). However, delayed feedback is outside of the scope of this dissertation. In addition, we acknowledge that, in this blended learning environment, students could get various other types of feedback from the instructor, TAs,

and other peer students in the classroom, in office hours, in problem-solving sessions, etc. These various types of feedback and myriad of other factors (e.g., assignment incentives, instructors, lecture structures, self-efficacy) all serve as important contextual factors that may moderate student behaviors and performance in this complex blended learning environment. As mentioned, for some of the important background factors for which we have a measure or a proxy measure in our quantitative data, we have included them as control variables or mediator variables in our model. Some of the factors are only represented in the qualitative interview and observation data, which we will consider as rich supplements in our discussion. There are other factors, which are not represented in our datasets or outside of the scope of this dissertation, we will discuss in our limitations section in CHAPTER 6.

3.2 Data Collection and Sampling Frame

As described in section 1.2, every first-year student in the studied institution has to take a Math diagnostic test. As a result, they were recommended to take the advanced, regular, or less difficult version of the Classical Mechanics course. A small number of students also choose to take the course in one of four first-year learning communities the university offers. Our study focuses on the mainstream version of the Classical Mechanics course—PHYS101, where a total of 474 students were enrolled during fall 2014. In some sense, the Math diagnostic test served as a sampling frame to exclude notable outliers from our sample. Our sample is thus representative of the majority of students who enrolled in the regular version of the course. We also utilized a stratified random sampling procedure when recruiting participants for interviews and observations, described in more detail below.

The following subsections detail all datasets we have collected for this study.

3.2.1 Server Tracking Logs

The edX course platform records students' every interaction with the course website as a browser-side event with additional server-side events. For example, when the students open the homework problem page, this initiates a browser-side "problem-show" event. In another example, when the students submit one solution to check its correctness, this results in a browser-side "problem-check" event. At the same time, the server needs to actually perform the checking and return information indicating whether the solution is correct or not, which is a server-side "problem-check" event. As the server successfully completes the checking, it initiates a "problem-graded" event to confirm the submitted solution has been checked and graded. In Figure 3.1, we showcase a synthetic piece of raw log data illustrating a server-side "problem-check" event in the raw JSON format. From this example, we can see that this piece of log data contains the student's username and IP address, the timestamp when the event happened, and detailed information about the answers the student submitted and the answers' correctness. In addition to problem interaction events, there are also other events such as navigational events, video interaction events, and textbook events, which we will use to understand students' behaviors.

All the students' interactions with the course website resulted in over 30 million events in total from September 1st, 2014 to December 31st, 2014. This is the primary dataset in this study.

```
{
   "username": "AAAAAA",
   "user_id": 9999999,
   "ip": "999.99.9.99",
   "time": "2014-03-03T16:19:05.584523+00:00",
   "problem_id": "i4x://edx/AN101/problem/a0effb954cca4759994f1ac9e9434bf4",

   "event_source": "server",
   "event_type": "problem_check",

   "event": {
      "answers": {
         "i4x-edx-AN101-problem-a0effb954cca4759994f1ac9e9434bf4_2_1": "yellow",
         "i4x-edx-AN101-problem-a0effb954cca4759994f1ac9e9434bf4_4_1": [
            "choice_0",
            "choice_2"
         ]
      },
      "attempts": 1,
      "correct_map": {
         "i4x-edx-AN101-problem-a0effb954cca4759994f1ac9e9434bf4_2_1": {
            "correctness": "incorrect"
         },
         "i4x-edx-AN101-problem-a0effb954cca4759994f1ac9e9434bf4_4_1": {
            "correctness": "correct"
         }
      },
      "grade": 2,
      "max_grade": 3,
   }
}
```


Figure 3.1 Example log data of a server-side “problem-check” event

3.2.2 Performance Data

We have obtained all students’ performance data in the course including grades for each online and written homework problem, grades for the final exam, and cumulative grades in the course. These performance data, primarily the final exam scores and cumulative grades, are used as outcome variables in our regression models. The cumulative grade, as an outcome metric, encompasses numerous components of a student’s effort and engagement with the class, rather than a proxy measure of conceptual understanding like the final exam score (Pintrich & de Groot, 1990). The Pearson correlational coefficient between the final exam score and cumulative grade is 0.836.

3.2.3 "Checkable Answers" Survey

During November 2014, we administrated a Qualtrics survey to PHYS101 students. The survey asked questions such as how students use the "checkable answers" features, how many times they check before giving up, whether they prefer to work alone or with others, what their attitudes are towards the PHYS101 course and its social component, how they perceive the relationship between PHYS101 and their future, and whether they have any previous experiences with blended learning classes. The full, executed survey is included in Appendix A.

In this study, we used three measures from the survey: students' self-efficacy to perform in this course, their perception of the utility of this course for their future, and the students' perception of the utility of the "checkable answers" feature. Specifically, the survey asked the students to rate, using a 1-7 Likert scale, 13 items about their perception of the helpfulness of the "checkable answers" feature. This includes whether the feature helped their learning by "helping to build my confidence", "contributing to my knowledge of the topics", "helping me check for errors", "making learning easier", "motivating me to find the right answer", and "reducing my misconceptions". Detail regarding all the 13 items is included in Appendix B. We constructed a scale out of the 13 items by calculating the average score of items a student responded to. We found the within-scale reliability for this new scale to be high (Cronbach $\alpha \sim = 0.98$). The measure for the students' self-efficacy to perform in this course was adapted from the motivation scales in the Motivated Strategies for Learning Questionnaire (MSLQ) (Pintrich et al., 1993). The within-scale reliability Cronbach alpha of this 8-item self-efficacy scale is about 0.95. The measure for students' perceived utility of this course for their future is

adapted from (Husman et al., 2004), and the Cronbach alpha is about 0.79. The detailed scale item breakdown for each measure is shown in Appendix B, and items with negative wording were appropriately reverse coded for calculations.

A total of 266 students responded to and completed the survey, resulting in a response rate of 56.12%. We performed one-way ANOVA tests (equivalent to two-sample t-test here) to examine whether there were systematic differences between students who did or did not complete the survey. The results are shown in Table 3.2. Students who completed the survey performed significantly better in terms of cumulative grades and final exam grades. This result indicates that there exist systematic differences between students who did or did not complete the survey; therefore, we need to consider the potential biases when using data from students who completed the survey.

Table 3.2 Performance differences between students who did or did not complete the "checkable answers" survey

	Students who completed the survey (N=266[a])	**Students who did not complete the survey (N=208[b])**	**P-value**
Mean Cumulative Grade (%)	81.120	78.360	<0.001**
Mean Final Exam Score (max points: 200)	140.989	133.406	0.005**

*p <0.05, ** p < 0.01

[a] Only N=264 students who completed the survey had valid final exam scores.

[b] Only N=207 students who did not complete the survey had valid final exam scores.

3.2.4 Background Factors

We obtained data on student background factors from the Office of Institutional Research and relevant administrative data from the Registrar. These factors include

demographics information such as gender, ethnicity, family income, and parents' education level, whether the students have participated in the pre-college bridge program, data that indicating students' Math level (e.g., Math courses taken concurrently with PHYS101), and data indicating students' prior ability of scientific reasoning (e.g., SAT II subject test Science scores).

3.2.5 Interviews

During the second week of November 2014, we conducted 18 semi-structured individual interviews with students who enrolled in PHYS101. We recruited the participants using a stratified random sampling method based on students' cumulative performance level in the course at that time—Low: [0, 80%), Medium: [80%, 90%), High: [90%, 100%]. We also oversampled students who participated in the pre-college bridge program, because, historically, a large number of students participating in the bridge program have been underrepresented minorities (URM) and also because the bridge program is relevant to the PHYS101 content and productive learning strategies. During the interview, we asked questions about the students' first year college experiences and their experiences in PHYS101 and using the edX course website, especially the "checkable answers" features. We also asked about their learning strategies and their perceptions of intelligence. Each interview lasted about 30-45 minutes. All interviews were audio recorded and transcribed into texts. Each participant was offered a $20 compensation for completing the interview. The semi-structured interview protocol is included in Appendix C.

3.2.6 Observations

Around the same time when we conducted the semi-structured student interviews, we also conducted a total of 17 observation sessions where students came in-person to our controlled lab space and solved online homework problems with the edX platform. We allowed the students to bring their colleagues or friends with whom they usually work on PHYS101 homework problems. Out of the 17 observation sessions, 5 of them were students solving problems collaboratively. These participants were also sampled using the same stratified random sampling procedure described above. Each observation session was recorded using two cameras—one front camera recording the students' facial expressions and other physical actions and one back camera recording students' click and mouse events on the computer screen. We asked the students to follow a think-aloud procedure during the observation session, that is, they were encouraged to speak aloud what they were thinking. Except for this think-aloud procedure, we asked the students to do the homework problems as much as they would normally do. Each observation took about 45 minutes to one hour. Each participant and each of their colleagues/friends, if they brought any, were offered a $20 compensation for completing the observation session. The think-aloud prompt is included in Appendix D.

3.3 Quantitative Modeling Framework: Statistics vs. Machine Learning

Data analyses in this dissertation are primarily quantitative and are supplemented by relevant qualitative insights. There exist arguably two quantitative methodological paradigms or cultures in educational studies. One is the more traditional quantitative methodological culture built off of probability and statistical inference. The other one is the newly emerging educational data mining (EDM) field where data mining and

machine learning techniques are used to model large-scale electronically recorded student data. These two methodological cultures are actually inherited from the two distinct cultures of statistics and machine learning. Machine learning has a deep root in statistics. However, it has expanded to include many new methods, as well as new terminologies and traditions.

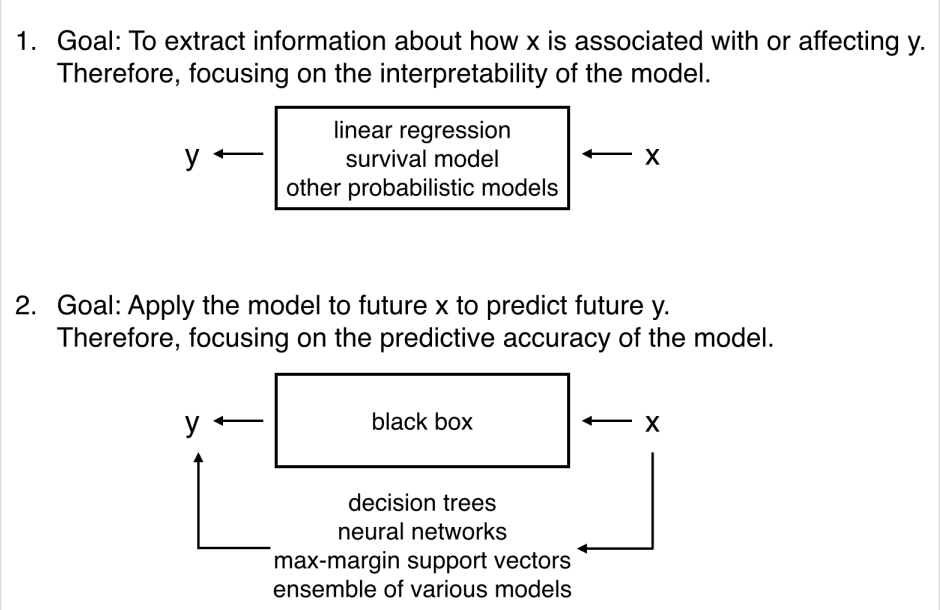


Figure 3.2 The two data modeling paradigms

At their essence, most data modeling methods try to use a vector of input variables $\boldsymbol{x}$ to model a response outcome variable $\boldsymbol{y}$. In statistics, the input variables are usually called independent variables, and the output variable is usually called the dependent variable. In machine learning and data mining, the input variables are usually called features, attributes, or dimensions, and the output variable is usually called the target value, or the class label in the case of a classification model. Breiman (2001) summarizes the two major goals of data modeling as shown in Figure 3.2, which in turn

lead to the two methodological cultures. Traditionally, most quantitative educational researchers are in the first culture group. However, as more and more online learning platforms and Massive Open Online Courses (MOOCs) are emerging, the amount of data available about students has exploded. More and more educational researchers start to adopt the methods from the second culture group.

The first goal is to extract information about how the input vector $\boldsymbol{x}$ is associated with the output variable $\boldsymbol{y}$ and how much effect each variable in vector $\boldsymbol{x}$ has on $\boldsymbol{y}$. Traditionally, most statisticians and quantitative social science researchers are in this culture group (Breiman, 2001). Statisticians call the modeling process "fitting", and the models are usually evaluated using goodness-of-fit or residual examination. This kind of model usually handles a smaller set of input variables compared with the second kind of model we will discuss. These input variables are usually theory-driven conceptual variables with operational definitions. Therefore, the modeling process is deductive in nature. One example of such model's application in educational studies is to model student performance using various factors and explain the effect of each factor (Aragon, Johnson, & Shaik, 2002; Powell, Conway, & Ross, 1990). Here, the primary goal is to estimate whether and how much the student performance is influenced by a particular factor or intervention strategy, rather than to use the model in a future time to predict student performance.

The second goal is to be able to apply the model in a future time on future input variables $\boldsymbol{x}$ to predict what the output variable $\boldsymbol{y}$ will be. Therefore, the focus is on how accurately the model can predict future $\boldsymbol{y}$, rather than whether the model actually reflects the process of generating $\boldsymbol{y}$ based on $\boldsymbol{x}$. Most machine learning researchers in Computer

Science are in this culture group. Computer scientists call the modeling process "learning" or "training". Models are usually evaluated by applying the model on a test dataset and comparing the predicted $\boldsymbol{y}$ of the test dataset and the "ground truth" $\boldsymbol{y}$ to obtain the predictive accuracy. To increase the predictive accuracy, computer scientists have developed various algorithms with complex structures that may not entirely based on probabilistic theory, including decision trees, neural networks, max-margin support vector machines, and ensemble of several models. These models usually perform better on much larger and more complex datasets than the type of models developed based on the first goal. In these models, there also exist ways to rank the relative importance of each input variable to the output variable. However, they often are not able to provide significance testing on treatment effects as thorough as statisticians do, and are more difficult to realistically interpret. This kind of model is usually applied to large-scale electronically recorded datasets and the modeling process is inductive in nature. In these applied contexts, it does not hurt to include a variable that does not contribute much to predict $\boldsymbol{y}$, because collecting data on this variable does not require extra work and including it does not hurt the model performance. Examples of applications of such models in educational studies include predicting student dropouts in MOOCs (Halawa, Greene, & Mitchell, 2014; Yang, Sinha, Adamson, & Rosé, 2013) and detecting potentially at-risk students based on their activities in online learning systems or social media sites (Baker, Corbett, & Koedinger, 2004; Chen, Vorvoreanu, & Madhavan, 2014).

Under the machine learning paradigm, when the "ground truth" $\boldsymbol{y}$ is not available, a strand of methods called unsupervised machine learning methods can be applied to discover patterns from the dataset inductively. The word "unsupervised" refers to the fact

that these methods do not rely on any known outcome variables to train the models. For example, Kizilcec, Piech, & Schneider (2013) used an unsupervised clustering method to identify four different types of student engagement patterns in MOOCs. Again, the purpose was not to predict the outcome variables in a future time but to discover patterns inductively.

In this dissertation, our primary goal involves understanding and interpretation of the influence of student behaviors on their course performance rather than accurate prediction of student performance in a future time, but we still borrow several methods from the machine learning culture, especially the concept of inductive quantitative analysis and unsupervised clustering methods. Overall, we adopt methodological traditions from both of the two quantitative data analysis cultures and use them in concert with each other. We explored and subsequently demonstrated how these two modeling approaches could inform each other. As illustrated in Figure 3.3, first, we extracted a set of behavioral variables from the tracking log data to describe students' interactions with the "checkable answers" and use of other online resources while solving the homework problems. The variable generation process was informed by a mix of both inductive and deductive methods: they were extracted in an effort to describe students' behaviors from various aspects as much as possible, but our reasoning for choosing these variables was still informed by our theoretical framework and many previous studies. Using this set of variables, we determined the cross-correlation between each behavior, including the hierarchical clustering of different behaviors. These clusters of related behaviors help us to better understand complex and interwoven usage behaviors. With this information and an application of the LASSO (Least Absolute Shrinkage and Selection Operator) variable

reduction method to multiple regression, we then estimated the impact of each remaining behavioral variable on the outcome, holding other behaviors constant. The hierachical clustering of behaviors help us identify groups of closed correlated variables, therefore to make conceptual justification of why certain variables were removed by the LASSO method. We went through a back-and-forth process and consulted the literature to make sure that variables having important theoretical groundings were not blindly removed by LASSO. This allows us to balance theory with our data needs. We clustered student behaviors to further determine overall patterns of usage and find meaningful groups of students with distinctive behaviors. We returned to traditional methods (t-test) to verify the significant difference in outcomes for these groups of students.

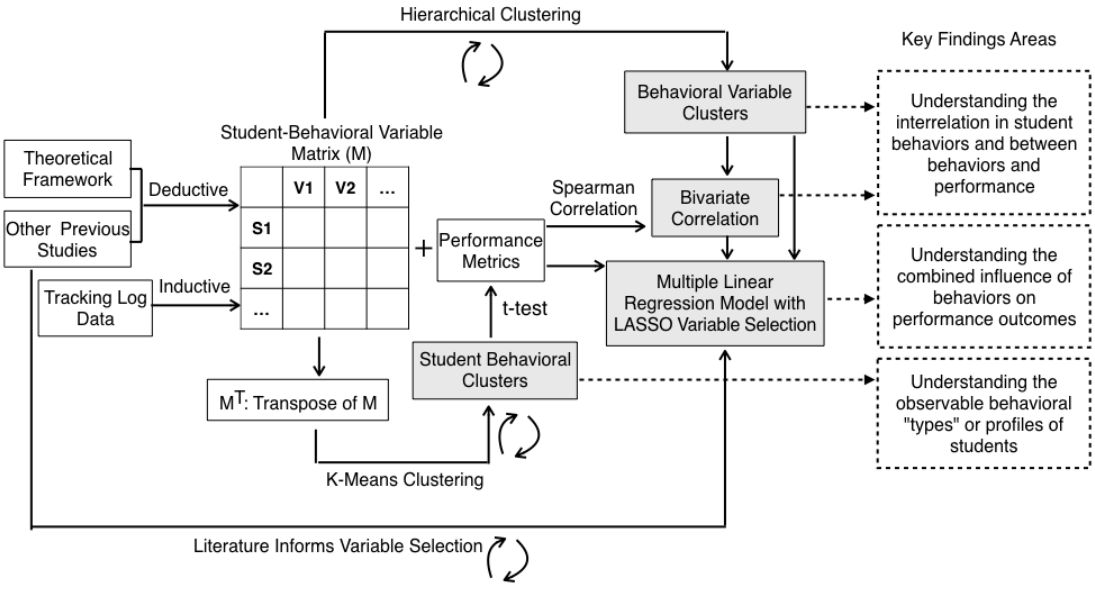


Figure 3.3 Quantitative data analysis diagram (findings are drawn from the four grayed areas)

In the following two subsections, we detail how we extracted the behavioral variables to describe students' behaviors with "checkable answers" during their problem-

solving process. We also delineate the unique challenges of missing data in this environment.

3.3.1 Extracting Behavioral Variables: Inductive vs. Deductive

We extracted 30 behavioral variables to describe students' behaviors while working on the online homework problems and 33 variables to describe their behaviors while working on the written homework problems. We first describe how we extracted the 30 variables for online homework problems, and then point out the few differences for the variables for written homework problems.

As mentioned, the variable generation process embraces a mix of inductive and deductive methods. Not all of the variables have clear theory-driven operational definitions, and they were generated in an effort to capture the variety of student behaviors we initially identified and uncovered over the process of conducting this study, without considering redundancy. Nevertheless, our reasoning for choosing these variables was informed by our theoretical framework initially and we check correspondence of our variables with many previous studies during the process.

To understand the student behaviors using "checkable answers", we start by looking at the basic patterns of checking. For each of the 22 online homework problems, a student submitted a set of checks, which we represent using a set of True or False symbols indicating the correctness. For example, if a student submitted solutions 5 times for a problem, and only the last solution was correct, then this checking result could be represented using the sequence [F, F, F, F, T]. Different students had different numbers of Fs and Ts on each homework problem, with different time intervals in between. Thus the first 15 variables were extracted based on students' sets of checks.

Then based on the students' time density of checks, we identified problem-solving sessions. If any two adjacent checks within a successive set of checks have less than 1 hour time interval in between, then we treat this as a dense cluster of checks. The time period starting at 30 minutes before the first check and ending at 30 minutes after the last check in this dense check cluster constitues a *problem-solving session* as shown in Figure 3.4. After identifying the problem-solving sessions, we could conveniently measure the frequency and duration of students' use of other online resources during the problem-solving sessions. The last 15 variables were thus extracted around students' problem-solving sessions and use of other online resources.

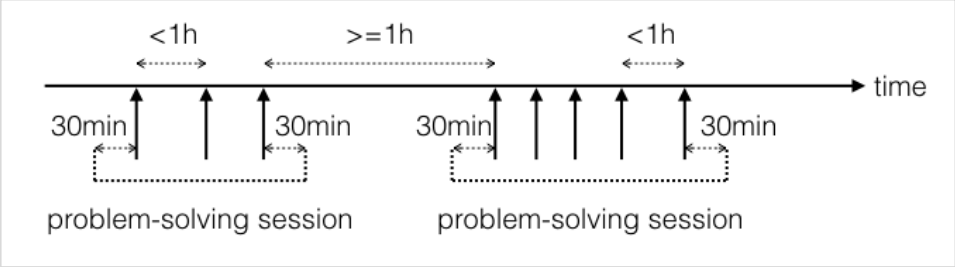


Figure 3.4 Problem-solving sessions (arrows are problem-check events)

We noted that different homework problems in this course had different difficulty levels. For some problems, most students could get to the correct answers within a few checks, while other problems took most students tens or even hundreds of checks. We used the total number of incorrect checks for each homework problem across all students to approximate the difficulty level of a given homework problem. For variables 5 and 6 below, whether the students got the correct answers on their first try or last try were initially binary descriptors of either 1 or 0 for each problem. We weighted these variables using the total number of incorrect checks for each homework problem across all students,

because we argue that if a student got the correct answer on their first try or worked until correct for a more difficult problem, it is reasonable to assign this student a higher variable value.

Table 3.3 contains the 30 variables and a brief description of how each variable is calculated. From an inductive variable generation perspective, we aim to use these variables to describe a variety of student behaviors without considering redundancy. More details on the range, mean, and median values for each variable are provided in Appendix E.

From a deductive variable generation perspective, several variables were initially identified based on our theoretical framework. For example, variables representing the average time period between checks or study sessions (variables 11, 12, and 20) could reflect whether students were taking the opportunity to reflect on the immediate feedback provided by "checkable answers", thus utilizing task level feedback to build from their current knowledge to increased understanding and potentially even transforming simple corrective feedback to higher level metacognitive skills. Or, if a student's number of checks was very high, and the intervals between checks was very short, that could mean that student was demonstrating relatively more hasty, instrumental, or careless checking behaviors.

Many other studies also informed our variable generation process. For example, researchers at the Carnegie Foundation for the Advancement of Teaching proposed the term "productive struggle" to describe the process where students struggle through difficulty and eventually achieve productive learning (Silva & White, 2013, p. 9). Therefore, effective measures for "struggle" could indicate engagement and persistence.

In our context, there were 10 weeks where the students were assigned two online homework problems. We used variables 13 and 14 to describe the time overlap between the two problems within each week, which could represent students having difficulty successfully solving the problems in order, therefore switching back and forth between the two assigned problems.

Table 3.3 Behavioral Variables for Online Homework Problems

Variable Name	Description
1 Correct checks	The average number of correct checks for each problem.
2 Incorrect checks	The average number of incorrect checks for each problem.
3 Total checks	The average of the total number of checks for each problem.
4 Correct fraction	The average fraction of correct checks for each attempted problem.
5 First correct fraction	The fraction of attempted problems where the student got it correct in their first try weighted by the difficulty level of the problem.
6 Last correct fraction	The fraction of attempted problems where the student got it correct in their last try weight by the difficulty level of the problem.
7 Not attempted	Number of not attempted problems.
8 First to due	The average time between the first check and the due time for each attempted problem.
9 Last to due	The average time between the last check and the due time for each attempted problem.
10 First to last	The average time between the first check and the last check for each attempted problem that have >= 2 checks.
11 First to second	The average time between the first check and the second check for each attempted problem that have >= 2 checks.
12 Interval between checks	The average time interal between all checks for each attempted problem.
13 Overlap time	The average overlapping time between the last check of the first problem and the first check of the second problem for each week where both two problems assigned in that week were attempted. If the last check of the first problem happened earlier than the first check of the second problem, then the overlap time is 0.

Table 3.3 continued

Variable Name	Description
14 Weeks overlap exists	The number of weeks where there was overlapping between the two online problems.
15 Activity after incorrect	The fraction of incorrect checks where other activities appear after the incorrect checks.
16 Num session	Number of problem-solving sessions.
17 Time all sessions	The total length of all problem-solving sesssions.
18 Avg session length	The average length of all problem-solving sessions.
19 Interval between sessions	The average time interval between problem-solving sessions.
20 Interval within sessions	The average time interval between checks within each problem-solving sesssion.
21 Video time	The average time spent viewing videos within each session. Based on industry accepted norm for timeout (Google, 2015; Interactive Advertising Bureau, 2009), if two successive activites appear within less than 30 minutes, then this time period in between the two activities is counted as spent on the first activity. The following variable 23, 25, 27 were also calculated based on this practice.
22 Video num	The average number of video clicks within each session.
23 Text time	The average time spent browsing the e-textbook within each session.
24 Text num	The average number of times that the student accessed the e-textbook within each session.
25 Class problem time	The average time spent checking the online PDF notes for in-class problems within each session.
26 Class problem num	The average number of times that the student accessed the PDF notes for in-class problems within each session.
27 Friday problem time	The average time spent checking the Friday problem-solving sesssion PDF notes within each session.
28 Friday problem num	The average number of times that the student accessed the Friday problem-solving sesssion PDF notes within each session.
29 Exam time	The average time spent checking the previous exam materials within each session.
30 Exam num	The average number of times that the student accessed the previous exam materials within each session.

Research related to choice-based assessments finds that students' actions preceding assessment activities can be as useful for assessing learning as the outcomes of the assessment (Schwartz & Arena, 2013). Some researchers also find that the greatest amount of student attention is focused on feedback to incorrect answers as opposed to feedback on correct answers (Timmers & Veldkamp, 2011). Grounded in this line of research, Krumm and colleagues used the percent of sessions wherein students attempted another activity after they gave an incorrect answer as one potential measure for "productive persistence" (Krumm et al., 2015). Based on this research, we included variable 15 – *the fraction of incorrect checks after which students attempted other activities*. This variable could represent that students took the effort to reflect on their incorrect answers and persist to find the correct answers.

In a study about student participation in an online course, Morris, Finnegan, & Wu (2005) conclude that engagement metrics—both in frequency and duration of participation—predict nearly one-third of the variability in achievement in an online course. Many studies on student behaviors in MOOCs and other online learning environments also usually measure student frequency and duration of using online resources such as video, e-textbook, etc. (Seaton, Bergner, Chuang, Mitros, & Pritchard, 2014; Seaton, Bergner, & Pritchard, 2013; Seaton, Nesterko, Mullaney, Reich, & Ho, 2014) We therefore included variables 21 to 30, which describe students' frequency and duration of using various online resources.

Following a similar procedure, we extracted 33 behavioral variables to describe students' behaviors while working on the written homework problems. However, there are two major differences compared with the variables for the online homework problems.

Table 3.4 Extra Behavioral Variables for Written Homework Problems

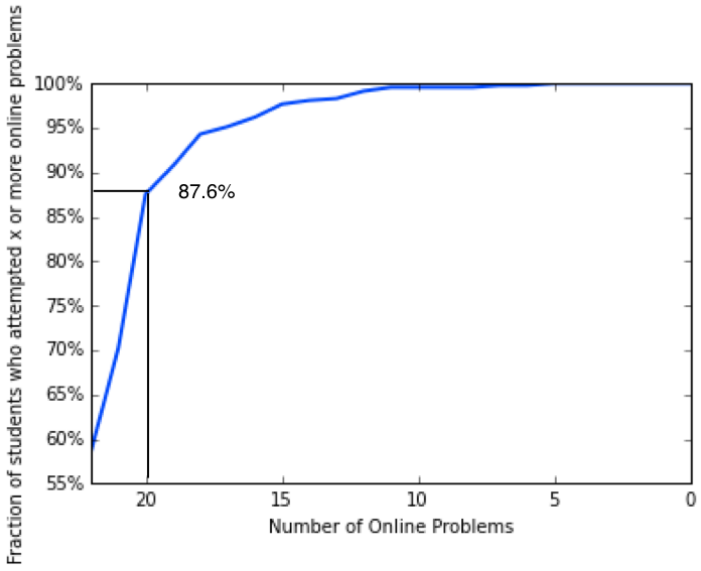
Variable Name	Description
1 Correct fraction before correct steps	For each step of a problem which is not the first step, for the first time the student got it correct, the fraction of previous correct steps. Take the average overall steps and all problems attempted.
2 Incorrect fraction before correct steps	For each step of a problem which is not the first step, for the first time the student got it correct, the fraction of previous incorrect steps. Take the average overall steps and all problems attempted.
3 Skipped fraction before correct steps	For each step of a problem which is not the first step, for the first time the student got it correct, the fraction of previously skipped steps. Take the average overall steps and all problems attempted.
4 Problems containing incorrect steps	The fraction of attempted problems where there were incorrect steps.
5 Problems containing skipped steps	The fraction of attempted problems where there were skipped steps.

First, there were only 1 or 2 online problems per week, but there were 5 or 6 written homework problems per week. So the calculation of overlapping time would be more complex. If we consider all possible permutations of 6 problems (6! = 720), the calculation becomes very expensive and inefficient for our goal. In fact, we will show in our data analysis results for the online homework problems, the measures of overlapping time did not turn to be effective measures for "productive struggle" in our context, therefore, we did not include such variables in the written homework analyses. Second, each written homework problem usually contains several steps. The immediate feedback for the written homework problems was stepwise feedback. To investigate whether students were taking advantage of information provided on previous steps to the problem-solving process and how those behaviors could influence their performance, we added the five variables in Table 3.4 to the behavioral variable set for written homework problems.

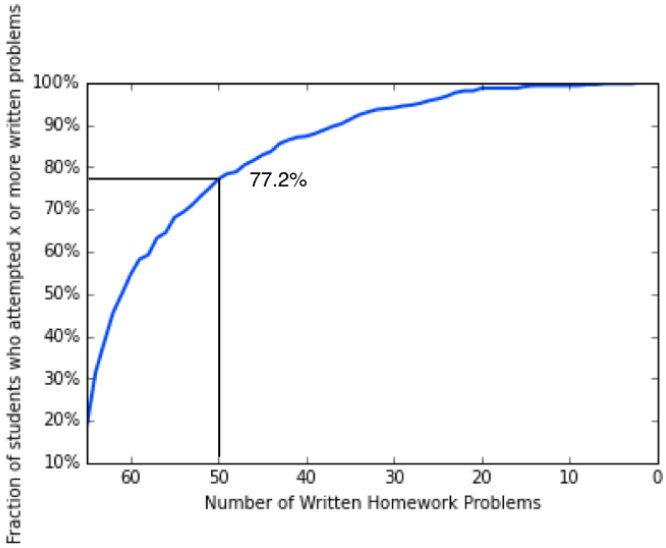
More details on the range, mean, and median values for each of these variables are provided in Appendix E.

3.3.2 Missing Data

Students use the online learning platform outside of the controlled classroom environment; therefore, they could use it in an infinite number of different ways or even not use it. In particular, checking the written homework problems online is optional in PHYS101. As it might be expected, in our dataset we see that not all students check every problem online, so we do not have an observation available for every student on every problem. Missing data is a pervasive issue in capturing student behaviors in online learning environment, especially in world-wide MOOCs, where some students skip several units of the course, or register but never interact with the course platform (Breslow et al., 2013; DeBoer, Ho, Stump, & Breslow, 2014; DeBoer, Stump, Seaton, & Breslow, 2013; Kizilcec et al., 2013). This is likely because MOOC students usually come from widely different backgrounds and have a variety of learning goals. In the blended learning context of PHYS101, all of the students were residential students, so we have less heterogeneity in the student sample, and missing data is less severe of an issue for our study. All students attempted some online homework problems, and only one student did not attempt any written homework problems. We removed this one student from our analysis of written homework problems. As shown in Figure 3.5, 87.6% students attempted 20 or more out of the 22 online homework problems. 77.2% students attempted 50 or more out of the 63 written homework problems. All students have some problem-solving sessions. The minimum number of problem-solving sessions for online homework is 5, and that for written homework is 3.



(a) Online homework



(b) Written homework

Figure 3.5 Complementary cumulative distribution function plots for students' number of attempted homework problems

Our decision as to how to deal with missing data is based on the mathematical nature of how we calculated each variable. Essentially, we have three types of variables.

The first type includes counts of checks (e.g., variables 1, 2, and 3), numbers of events (variables 22, 24, 26, 28, and 30), and time spent viewing certain resources (variables 21, 23, 25, 27, and 29). For variable type, we argue that if the student did not attempt a problem or did not use certain resources, this does not mean that we miss the data to capture the student's behaviors, but, rather, it means that not attempting the problem or not using the resources itself could be a conscious choice of the student and is itself a behavioral data point we should capture. Therefore, for this type of variables, we set the default value as 0. For the same reason, we included variable 7—*number of not attempted problems* in our variable sets. We also included variables to capture students' skipped steps in the variable set for the written homework problems. Setting the default value to 0 is the unbiased way to capture the student's behavior of not attempting a problem or not using certain resources. The drawback of doing this is that we miss the opportunity to estimate the student' latent ability on the not attempted problems.

The calculation of the second type of variables involves fractions (e.g., variables 4, 5, 6, and the five extra variables for the written homework). Use variable 4 (*the average fraction of correct checks*) as an example, for not attempted problems, it is not justified to set the default of this variable as 0, because we do not know how many checks would be correct if the student did the problem, setting the fraction to 0 would bias the average value downward. Therefore, for this type of variables, we take the average of attempted problems.

The third type of variables calculates the interval between two events (e.g., variables 8, 9, 10, 11, 12, and 13). For this type of variables, it is also not justified to set the default as 0, because that would bias the time interval between two events towards a

shorter time, and introduce the illusion that the student was interacting with the problem with a higher frequency than was actually true, while actually they did not even attempt the problem. Therefore, we deal with this type of missingness the same way as for the second type of variables—we used the student's average behaviors across all problems that they actually attempted. This way of dealing with missingness essentially uses the student average behavior on attempted problems to estimate the student's latent ability on the not attempted problems. As we will show in Figure 4.5, different problems have different difficulty levels, and thus, students may take different numbers of checks to get to the correct answers, and demonstrate different behaviors on different problems. For this reason, our way of dealing with missingness here may introduce higher variance in the estimates due to problem differences.

In the following three subsections, we give brief descriptions of the three modeling methods we use in this dissertation: agglomerative hierarchical clustering, multiple linear regression with LASSO (Least Absolute Shrinkage and Selection Operator), and K-means clustering method.

3.3.3 Agglomerative Hierarchical Clustering

We conducted Spearman correlation analysis between each of the behavioral variables and the students' performance metrics and self-efficacy and perception measures. We chose to use Spearman rank correlation instead of Pearson correlation because, as we will demonstrate in CHAPTER 4, the distributions of most variables among all students are highly skewed.

One characteristic of the dataset that was immediately apparent was that the behavioral variables were not completely mutually independent. This is quite common in

research on human behaviors and other MOOC studies on student behaviors (DeBoer et al., 2014). Student behaviors are mediated by various background factors such as prior knowledge, computer proficiency, metacognitive skills, and motivation etc. Therefore, two or more behavioral variables are likely mediated by a common set of latent traits and thus could be highly correlated with each other. It is unsurprising that some variables, such as the number of lecture video accesses and the time spent viewing lecture videos, are almost necessarily correlated with each other.

To understand the correlation among variables, we not only present the correlation matrix among the variables, but also perform agglomerative hierarchical clustering on the 30 variables for online homework problems and 33 variables for written homework problems respectively. Agglomerative hierarchical clustering, by its name, is a bottom-up clustering method (Rokach, 2010). Each cluster, containing only one variable at the beginning of the clustering process, is merged with another cluster having the shortest distance as the hierarchy moves up. The distance measure between two clusters used in this study is the average Spearman correlation distance. The more positively correlated two variables are, the shorter their distance is. Similarly, the more negatively correlated the two variables are, the longer their distance is. This clustering algorithm runs recursively until all the variables are grouped together, and we thus get a dendrogram structure of the variables, which helps us understand the correlation among them.

3.3.4 Multiple Linear Regression with LASSO

To understand the amount of variance in students' course performance explained by students' behaviors, we performed multiple linear regression analysis. All behavioral variables were normalized to values between 0 and 1. We included the students' background factors such as gender, ethnicity, whether they have participated in the pre-college bridge program, need-based financial aid information, concurrent Math courses, and SAT/ACT scores.

The multiple linear regression model can be written as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

The coefficients are calculated using least squares method, therefore, this method is also called Linear Least Squares regression or Ordinary Least Squares (OLS) regression. The multiple regression model is easy to understand and highly interpretable. The coefficients can be used to interpret the relative importance of each variable to the output variable $\boldsymbol{y}$, holding all other variables constant.

However, we often cannot guarantee that all input variables are uncorrelated, especially when the variable generation process is entirely or partially inductive and we do not have a thorough theoretical understanding of every single variable. Many of the behavioral variables were correlated with each other, so we had a low level of tolerance for many of the predictors. When we included all behavioral variables in the regression model, the variance inflation factors (VIF) for these variables ranged from 1.26 to 612.74 (VIF=1 if none of the variance in that variable is shared with other regressors). Although this does not introduce bias, would decrease reliability of point estimates.

From the machine learning perspective, one possible solution is to introduce higher order terms such as x_1x_2 into the model to indicate the interaction between two variables. If we have a large enough dataset, which is the ideal application context of a machine learning algorithm in the big data industry, we can include as many as possible higher order terms, without influencing the model performance much. However, the practical fact is that we have a decently but not infinitely large dataset. With more than 30 variables, the possible number of higher order terms increase exponentially. As the model includes more and more higher order terms, the computation gets more and more complex. There exits an upper limit to the model complexity with any acceptable degree of freedom, that is, the model could suffer from a degree of freedom constraint problem (Babyak, 2004). In the machine learning culture, this concept is explained as an overfitting problem, that is, as the model complexity increases, the model performs better on the training dataset, while worse on the testing dataset. Further more, introducing nonlinear terms make the model difficult to interpret, which deviates from our goal of understanding the effect of the behavioral variables on the outcome variables.

From a traditional statistical perspective, one possible explanation to the inter-correlation among large number of variables is that they can be modeled as linear combinations of a smaller set of unobserved variables. This is the intuition from factor analysis. The reasons we did not use factor analysis eventually are (1) we did not have a clear theoretical understanding of how many and what these unobserved variables mediating the observed behavioral variables are, and it is not our goal to identify these unobserved variables; (2) without clear theoretical explanation of the unobserved variables, the results of factor analysis are hard to interpret and could not lead to a fully

specified regression model. Nevertheless, the hierarchical clustering analysis of the behavioral variables still drew upon the intuition of factor analysis. Closed correlated variables in the same cluster are likely to be mediated by similar set of unobserved variables.

We eventually chose to use the LASSO (Least Absolute Shrinkage and Selection Operator) method to reduce the number of variables, because LASSO tends to pick one out of a group of correlated predictors and discard the others, and reducing the number of variables can avoid the overfitting problem as well (Hastie, Tibshirani, Friedman, & Franklin, 2013). LASSO adds a penalty term to the loss function of the regression, which is the sum of the absolute values of the coefficient estimates multiplied by a parameter λ. The larger λ is, the more coefficients are shrunk to zeros, thus serves the purpose of variable selection. The penalty term added is also called L1 regularizer. The main purpose of using regularization is to control the model complexity and avoid model overfitting. In the case of multiple linear regression, we aim to use the least amount of variables that comprise a fully specified model and explain a similar amount of variance in the dependent variable. We balance our need for a fully specified model that captures meaningful predictors of performance with the recognition that some of the variables we have extracted may be nearly representing the same behaviors – LASSO is useful for our analyses by helping select predictors (Hastie et al., 2013).

After applying the LASSO variable selection, we then run OLS multiple linear regression using students' performance measures on the selected variables. We perform F-tests to compare the regression models before and after the variable reduction to test whether the reduction of variables significantly changed the amount of variance

explained in the students' performance metrics. We used the residual plot and Cook's distance to identify and remove outliers. Cook's distance measures the effect of removing a data point from the regression model. Data points with large Cook's distance can be seen as outliers that distort the outcome. Then we used the Shapiro-Wilk normality test to verify that the residuals were normally distributed.

3.3.5 K-Means Clustering

In order to identify clusters of students who demonstrate vastly different behavioral patterns, we applied the K-means clustering method. As opposed to the hierarchical clustering method, K-means clustering is a popular partition-based method (Hartigan & Wong, 1979; Rokach, 2010). It is a heuristic clustering method, but it has been widely demonstrated to be very effective in various application contexts (Bishop, 2007).

In standard K-mean method, the algorithm first randomly selects $\boldsymbol{k}$ data samples as centroids. The rest of the samples are put into the same cluster as the centroid to which they have the shortest distance. The distance between any sample and the centroid of a cluster used here is the Euclidean distance. The centroid of each cluster is then recalculated based on the mean of all samples in that cluster. This process is iterated until the change of the centroid is smaller than a certain threshold. The choice of the number of clusters $\boldsymbol{k}$ is usually based on domain knowledge, or to first determine a possible range of $\boldsymbol{k}$ based on domain knowledge, and then exhaust all $\boldsymbol{k}$ in that range to find the best one.

Our goal here is to identify the various student clusters and thereby compare their performance metrics, rather than to obtain a hierarchy of relationships as in section 3.3.3. The unit of analysis here is the student, each student with a set of attributes, where in our

agglomerative hierarchical clustering (section 3.3.3), the unit of analysis is the behavioral variable, each variable with 474 values from all students. This can be effectively understood as the input student-variable matrices being transposes of each other in these two clustering analyses.

There are certain limitations with the K-means method. For example, the choice of the number of clusters $\boldsymbol{k}$ is usually largely based on domain knowledge. Therefore, in less well-understood research contexts like ours, we need to go through an iterative process and triangulate with other data analysis results in order to find the value of $\boldsymbol{k}$ that is most meaningful for educational practices.

We present the results of our data analyses using these methods in the next chapter.

CHAPTER 4. RESULTS

This research answers three major research questions. The first two research questions focus on student behaviors with "checkable answers" and other online resources while working on the online homework problems. The third research question digs into the differences in student behaviors for the written homework compared with that for the online homework. The following section 4.1 presents results to answer the first two research questions, and section 4.2 delineates the differences of student behaviors in the two different assessment tasks.

4.1 Results for Online Homework

The "checkable answers" for online homework problems provided students task-level immediate corrective feedback. Results of our analyses indicate that there are a wide variety of student behaviors; many, but not all, are significantly related to performance and attitudinal outcomes, though only a few are strongly so.

For all the 22 online homework problems, there are a total of 58,422 server-side "problem-check" events from the 474 students. This means that in total, students in PHYS101 have performed 58,422 problem checks with "checkable answers" for online homework. The total number of correct checks is 9,545, and that of incorrect checks is 48,877, which means only 16.3% of the checks resulted in a green check mark. The average number of checks for each student is 123 and the median is 109. The maximum

number of checks for a student is 821, and the minimum is 21. Students are indeed utilizing the "checkable answers" feature as a resource, on average, checking each problem about 5 times. Figure 4.1 is a stacked bar graph illustrating the number of correct and incorrect checks for each student (ordered based on total number of checks descending).

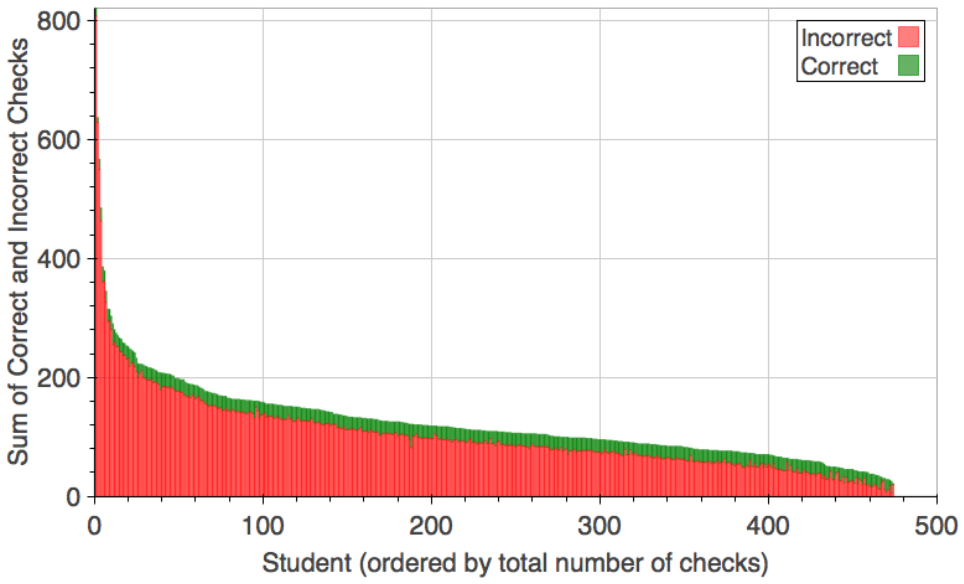


Figure 4.1 Number of correct and incorrect checks for each student over all 22 online homework problems

4.1.1 Skewness

The distributions of most of the variables are highly skewed, hence the rationale for our choice to use Spearman rank instead of Pearson correlation. Although these behaviors demonstrate a high level of skew, they are not as notably skewed or widely varied as behaviors in the solely online MOOC context (DeBoer et al., 2014). We highlight the following 3 variables to demonstrate this point. In the following three

density plots, the area under curve up to a certain ***x*** is the probability that the ***y*** <= ***x***, and each dot on the **y** = 0 axis represents a student at a specific ***x*** value.

From the density plot in Figure 4.2, we can see that the distribution of total number of online homework checks for each student is skewed, with a large number of students performing around 100 checks and a few students performing more than 300 checks. We can observe that, if we roughly remove the outlier students who performed more than 300 checks, the distribution approaches a normal distribution. We see that many behaviors can be characterized as a few outliers demonstrating the extremes of this behavior, with the bulk of the students demonstrating an "in-between" level of this behavior.

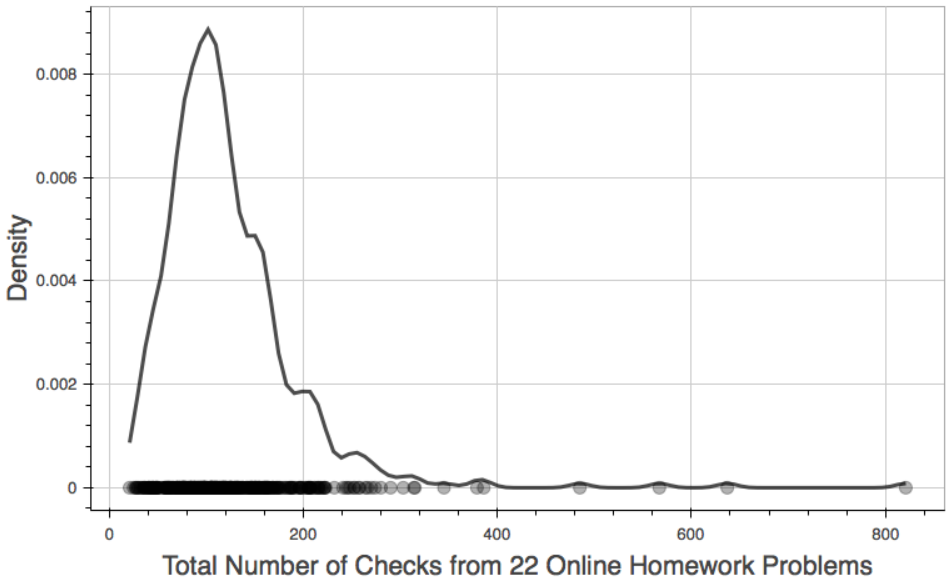


Figure 4.2 Density plot of total number of checks from 22 online homework problems for each student (variable 3)

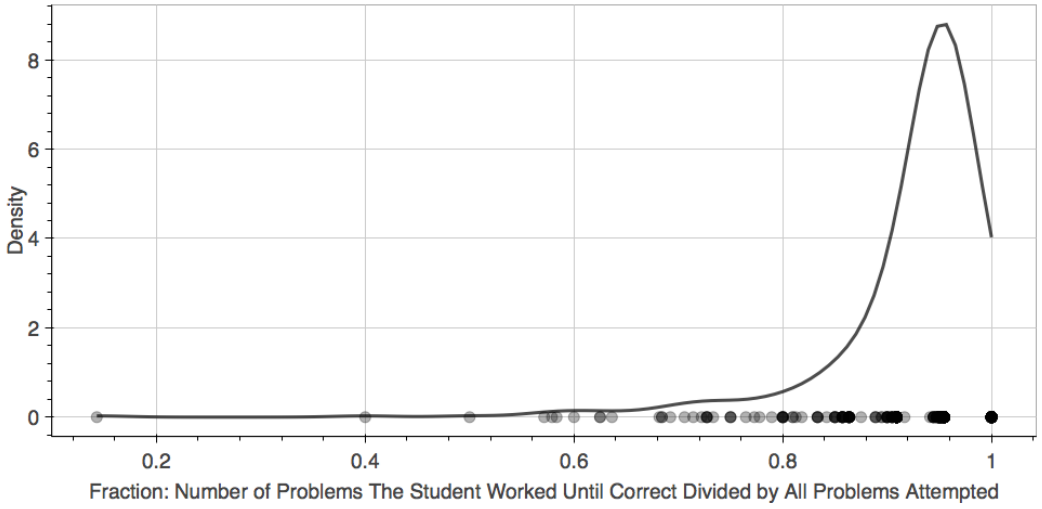


Figure 4.3 Density plot for the fraction of problems where the student worked until correct over all problems attempted (variable 6 before weighting using difficulty levels)

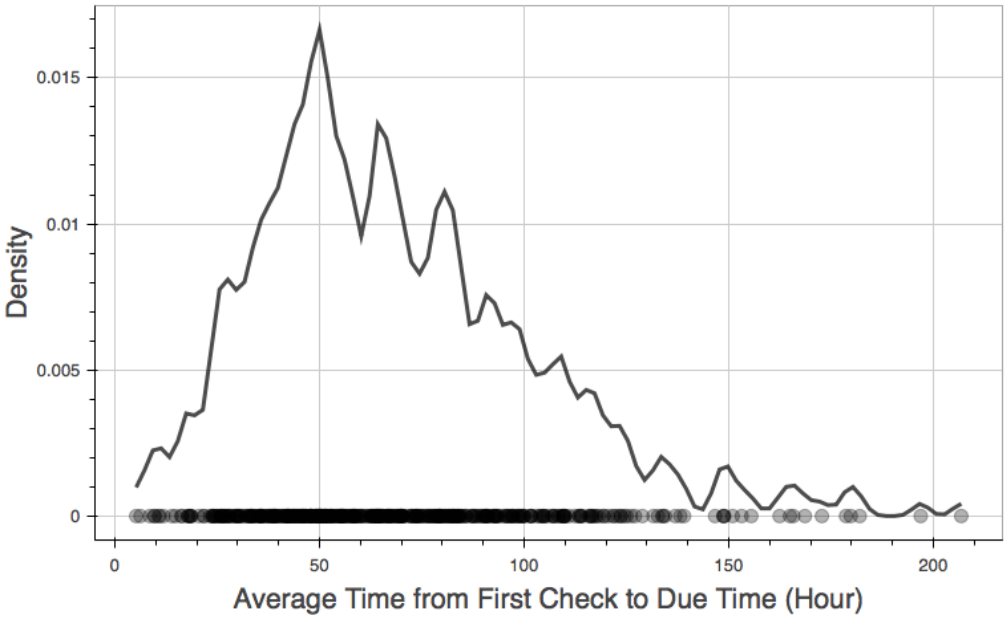


Figure 4.4 Density plot for the average time across all homework from the first check to the homework due time for a given student (variable 8)

4.1.2 Difficulty Levels

The difficulty levels of the 22 online homework problems vary. We use the total number of incorrect checks to approximate the difficulty level of each homework problem. The minimum number of incorrect checks for a homework problem averaged across all students is 0.27, while the maximum number is 11.92. In Figure 4.5, the ***y***-axis shows the total number of incorrect checks for each homework problem, and the ***x***-axis is the homework due time. The dates for the 3 mid-term exams and final exam are also marked. Problems in the same week were connected together. We can see that the difficulty of problems was often lowered before exam time.

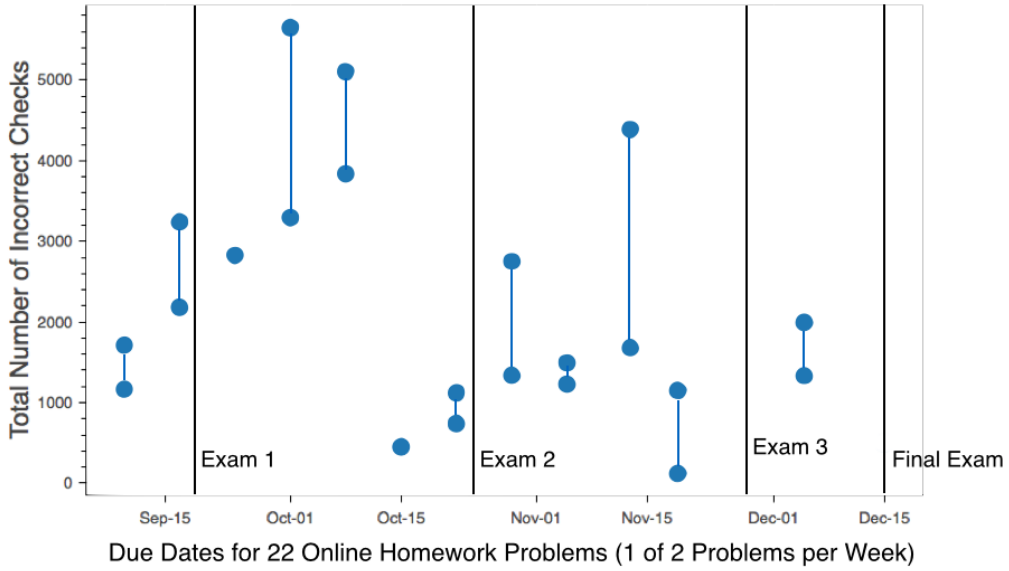


Figure 4.5 Number of incorrect checks on each homework problem which we use to approximate difficulty levels

4.1.3 Behavioral Variables Clusters and Correlation Analysis

Figure 4.6 shows the dendrogram structure of the 30 behavioral variables for online homework problems based on their Spearman correlation distances. The 5 basic

variable clusters are annotated and segmented using dashed lines. Figure 4.7 shows the Spearman correlation coefficients between each variable and students' performance, self-efficacy, and attitudinal measures. The variables are ordered based on the hierachy in Figure 4.6, and different variable clusters are segmented by black lines. Variables in the first two clusters are mostly positively correlated with students' performance measures, while variables in the fifth cluster are negatively correlated with students' performance measures. Variables in the third and fourth clusters are not significantly correlated with students' performance measures. The correlation matrix of pair-wise correlation among all 30 variables is given in Appendix F.

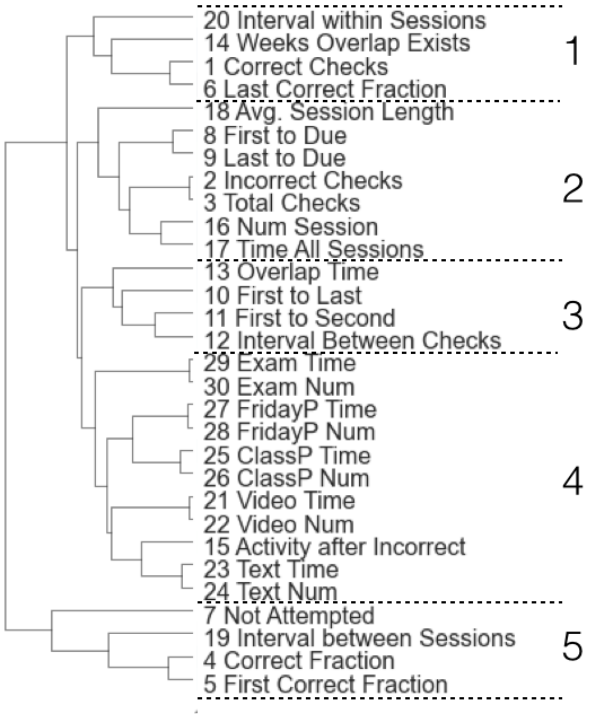
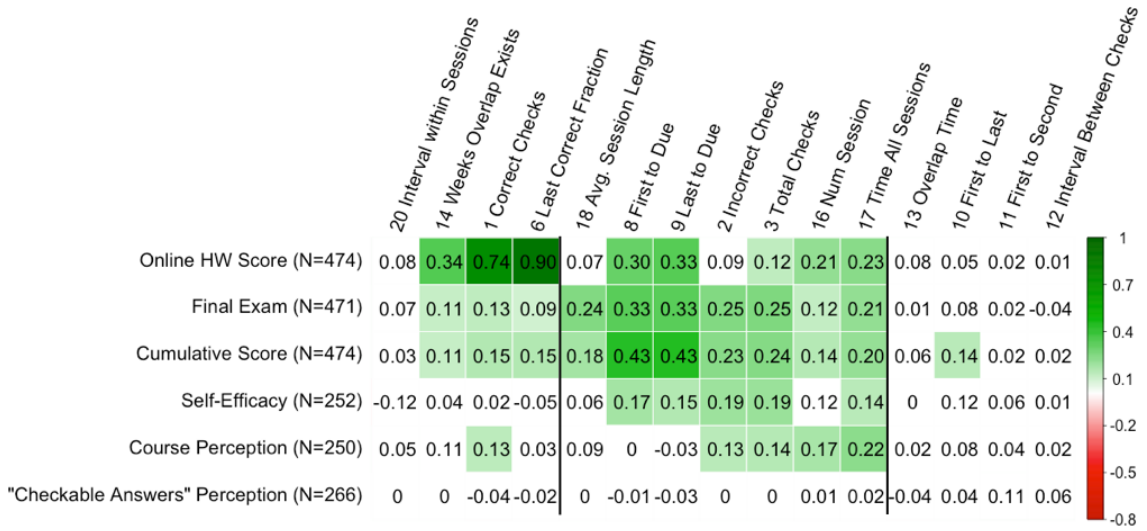
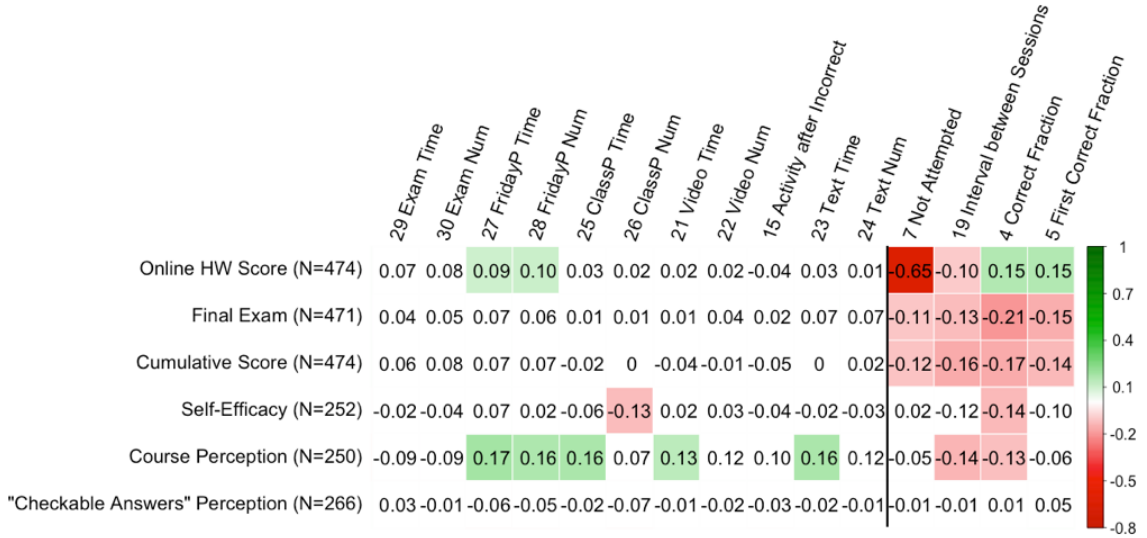


Figure 4.6 Hierarchy of the 30 behavioral variables for online homework based on their Spearman correlation distances



(a)



(b)

Figure 4.7 Spearman correlation coefficients between the 30 behavioral variables for online homework and student performance and attitudinal scores from the survey (cell background is white if p >= 0.05)

By looking at the dendrogram structure in Figure 4.6, we can easily understand the correlation between variables. Highly positively correlated variables were merged

together first. In the first cluster, variable 1 is the number of correct checks, and variable 6 is the weighted fraction of problems attempted where the students worked until correct. For problems where the students worked until correct, 97.3% only had 1 correct check (a near-constant). Therefore, variable 6 is approximately the problems where students worked until correct divided by the total number of problems attempted (87.6% students attempted 20 or more problems, also a near-constant), while variable 1 is approximately the problems where students worked until correct multiplied by a near-constant. Essentially, both variables describe students' answer-until-correct behavior. Variables 13 and 14 were designed to measure the students' "productive struggle", but these two variables were not merged together by the hierarchical clustering method. In fact, variable 14 was merged together with variables 1 and 6. This means that the number of weeks where students demonstrated back-and-forth behaviors correlates with the number of problems they checked until correct. Variable 13 measured the length of the overlapping time, and did not turn out to be significantly correlated with any of the performance measures. Thus variable 13 is either not an effective measure for "productive struggle", or this type of behavior is not productive for learning for students in our sample.

In the second cluster, variables 8 and 9 are highly positively correlated. This means that students who start the homework early (have longer first check to due time) also usually submit the homework early (have longer last check to due time). This behavior positively contributes to high performance in the course. Variables 2 and 3 are also highly positively correlated. This is obvious because the total number of checks is the number of incorrect checks plus a near-constant number of correct checks. What is

surprising is that both of these two variables positively contribute to students' performance, while in the fifth clusters, variable 4 (fraction of correct checks) and variable 5 (fraction of problems where students got the correct answers on their first checks) are negatively correlated with students' performance measures. Put simply, this implies that it does not matter whether the students got a lot of incorrect answers, as long as they kept trying and kept engaging with the feedback and problems until they got the correct answers. We have hypothesized that getting the answers correct on the first check might mean that the students had higher prior knowledge. However, our findings did not show significant evidence for this assertion. As we show in Figure 4.8, students who have a high first correct fraction are those who submit the homework late (within 2 days of the due date). Because variable 8 and variable 9 are highly correlated, students who submitted the homework late were also students who started the homework late. These students have a high fraction of correct checks, so they necessarily have high online homework scores. However, their behaviors seem to actually hurt their performance as measured by final exam scores and cumulative grades. We could imply that these students might be those who were copying answers from other students at the last minute, rather than those who had high prior knowledge on the subject. At minimum, these students had very little engagement with the material and the immediate feedback.

Variables 16, 17, and 18 in the second cluster are all positively relate to students' performance. In other words, students measured with longer, more concentrated problem-solving sessions tend to perform better.

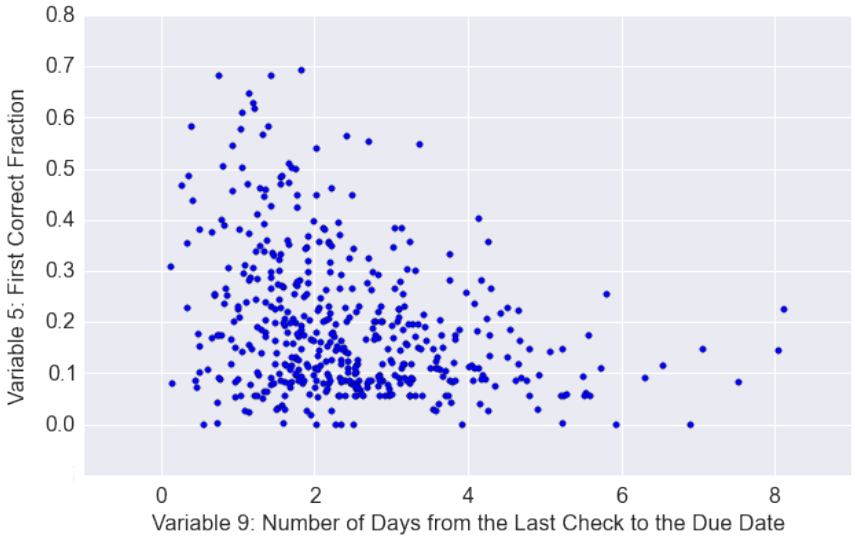


Figure 4.8 Students who submit the homework late have higher first correct fraction

Variables in the third and fourth variable clusters do not have a significant relationship with student performance. This may indicate that use of other online resources in the course platform is not a good measure for engagement, because in this blended learning context, students have many other resources such as peer discussion, tutoring sessions, and in-class problem-solving sessions to obtain help, rather than using the online resources. This reflects one of the major differences between blended learning courses and entirely online courses.

In the fifth cluster, variable 7 is the number of not attempted online homework problems. Despite the fact that the online homework scores only count for 2% of the cumulative grade, not doing the online homework was still negatively related to students' final exam performance and cumulative grades. Variable 19 is the average time interval between problem-solving sessions. We can see that if the students were inactive on the

platform for too long in between problem-solving sesssions, that was negatively related to their course performance.

Self-efficacy measures are positively correlated with most variables in the second cluster. Students' perceived utility of the course is positively correlated with most variables in the second and fourth clusters. Surprisingly, students' perceived utility of the "checkable answers" is not significantly correlated with any of their behaviors.

4.1.4 Multiple Linear Regression with LASSO

We first performed OLS multiple linear regression using all the behavioral variables and control variables with the dependent variable as the cumulative grade. We first included all data samples in the model, and did Shapiro-Wilk normality test on the residuals. The result ($w = 0.978$, $p < 0.001$) indicated that the residual distribution was significantly different from normal distribution. So we used the Cook's distances to identify and remove outliers. After removing six most deviant outliers, the Shapiro-Wilk test result was $w = 0.995$, $p=0.100$, which means under a significant level of 5%, there was no significant difference between the residual distribution and normal distribution. Figure 4.9 shows the Cook's distance plots and normal Q-Q plots before and after removing the outliers. We can see that the range of Cook's distances is much smaller after removing the outliers than before.

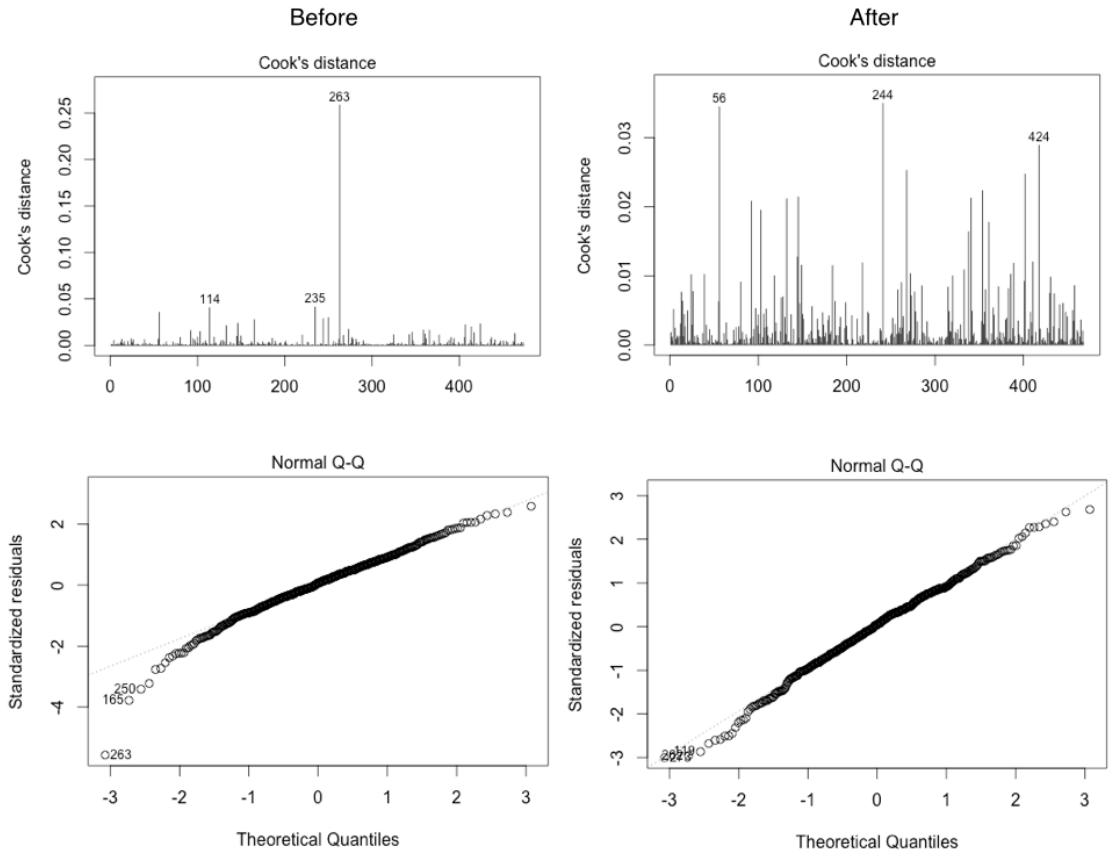


Figure 4.9 Cook's distance plots and Q-Q plots before and after removing outliers

We have noted that including all variables in the regression model resulted in high variance inflation factors (VIF) ranging from 1.26 to 612.74. This could be due to the fact that many variables were likely referring to closely related behaviors or were mediated by similar background factors. Therefore, for each pair of variables with |correlation coefficients| > 0.85, we removed one of the variables from that pair. There were 20 variables left. Though not strongly so, many of the rest of the behavioral variables are still correlated with each other, which would still inflate the variance of the coefficient estimates in regression models. We therefore used LASSO to do variable selection in order to control model complexity and remove some of the more inter-correlated

variables. We chose the λ parameter for LASSO when the mean-square error is minimized as shown in Figure 4.10

Based on the LASSO variable selection result and our theoretical framework on feedback effects, we selected the variables to be used in the regression model. The second column of Table 4.1 shows the regression results using the selected 11 behavioral variables as regressors for students' cumulative grades. The variance inflation factors (VIF) for selected variables range from 1.10 to 1.76, which means the influence on the standard error by the correlation among variables is at reasonable levels. The third column of Table 4.1 shows the coefficient estimates when include all the control variables such as gender, ethnicity, financial aid status, and Math level. Including these control variables has slightly changed the coefficient estimates. In general, the standard errors for the estimates became smaller. For example, the coefficient for variable 8 changed from 19.349 to 18.714, and the standard error reduced from 2.248 to 2.274. Before including the control variables, some variability of the outcome variable is explained by the error term in the model. Adding the control variables reduced the bias of the coefficient estimates. It is worth noting that, after including the control variables, the coefficient estimate for variable 6 – *last correct fraction* increased from 4.179 to 5.194 and the standard error has become smaller. This indicates that the estimate for this coefficient might be suffered from attenuation bias, and including the control variables served to correct this bias.

We performed an ANOVA test to compare the models before and after the variable selection ($F = 0.77$, $p = 0.75$), finding no significant difference in the amount of

variance explained before and after the variable selection despite removing 19 out of the 30 behavioral variables.

Following the same procedure, the fourth and fifth columns of Table 4.1 shows the regression results after variable selection when using final exam grade as the dependent variable. Even fewer variables were selected when the dependent variable is final exam score. Approximately 25% of the variance in students' cumulative performance and 16% of variance in students' final exam scores can be explained by students' behaviors with the online homework problems.

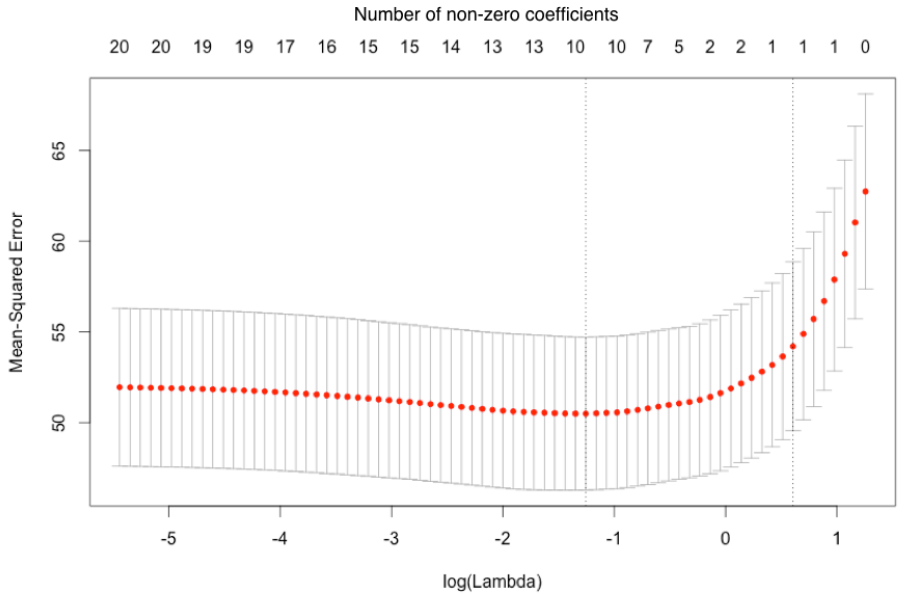


Figure 4.10 Mean-squared error with error bar from 10-fold cross-validation at each value of λ using LASSO when dependent variable is cumulative grade

Table 4.1 Regression Model Using Variables Selected by LASSO

Variable	(1) (Dependent Variable: Cumulative Grade: Max 100)	(2) (Dependent Variable: Cumulative Grade: Max 100)	(3) (Dependent Variable: Final Exam: Max 200)	(4) (Dependent Variable: Final Exam: Max 200)
Weeks Overlap Exists (v14)	3.057 (Std. error: 2.190)	2.182 (2.157)	10.310 (7.509)	9.171 (7.304)
Last Correct Fraction (v6)	4.179 (3.317)	5.194 (3.236)	--	--
Avg. Session Length (v18)	4.191 (2.300)	3.709 (2.274)	30.435** (8.021)	24.400** (7.859)
First to Due (v8)	19.349 ** (2.248)	18.714** (2.229)	50.915** (7.823)	51.150** (7.729)
Incorrect Checks (v2)	-6.896 (4.485)	-6.491 (4.453)	-18.661 (15.913)	-17.383 (15.605)
Overlap Time (v13)	-0.045 (2.361)	0.746 (2.298)	3.087 (8.822)	7.181 (8.502)
Exam Time (v29)	-5.202 (3.287)	-5.528 (3.217)	-24.812 (13.705)	-20.768 (13.242)
Video Time (v21)	-3.317 (2.109)	-2.607 (2.095)	--	
Activity after Incorrect (v15)	-1.812 (2.422)	-2.918 (2.360)	--	
Text Time (v23)	-2.196 (2.580)	-1.552 (2.531)	--	
First Correct Fraction (v5)	-1.502 (2.107)	-1.592 (2.059)	--	
(Intercept)	69.266 ** (2.763)	63.657** (3.151)	108.724** (5.852)	93.057** (8.318)
	Residual std. error: 6.589, R^2: 0.252, p < 0.001** No controls	Residual std. error: 6.342, R^2: 0.316, p < 0.001** With controls	Residual std. error: 24.690, R^2: 0.163, p <0.001** No controls	Residual std. error: 23.440, R^2: 0.269, p <0.001** With controls

Control variables: gender, ethnicity, financial aid, first-generation college student, pre-college bridge program, concurrent Math courses, and SAT II subject test Science scores

4.1.5 Two Student Clusters

To identify student clusters with various behavioral patterns, we performed K-means clustering. We went through an iterative process to choose the number of clusters. We tested for 2, 3, 4, and 5 clusters, and found that when the number of clusters was 2, the results corresponded with our findings in previous sections and made meanful and practical sense. The two student clusters demonstrated distinctive behaviors. Figure 4.11 compares the centroids of the two clusters. Centroid behavior is not the behavior of one specific student; rather, it can be seen as the average behavior of all students in that cluster. Cluster A contains 256 students, while cluster B contains 218 students. Students in cluster A demonstrated all the more productive behaviors and performed significantly better compared with students in cluster B in both cumulative scores and final exam scores (t-test with $p < 0.001$). The variables in Figure 4.11 are ordered and shaded based on the hierachy in Figure 4.6.

Using the multiple regression built in the last section, the predicted cumulative grade for the centroid of cluster A is 81.97, and that of cluster B is 77.92. The predicted final exam grade for the centroid of cluster A is 143.86, and that of cluster B is 130.97. This results verify the effectiveness of our regression model.

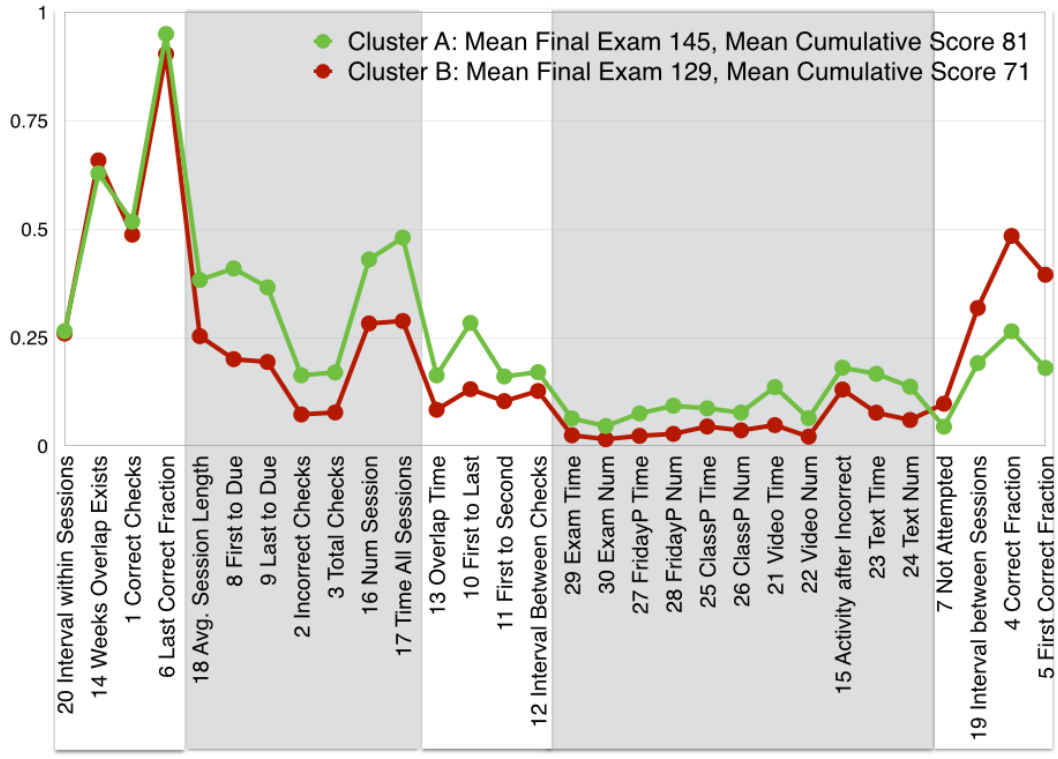


Figure 4.11 Comparison of the centroids of the two student clusters for their behaviors on online homework problems (all variables normalized to 0-1)

4.2 Results for Written Homework

We generated 33 behavioral variables to describe student behaviors while working on the written homework problems, and followed similar procedure as for the online homework to analyze these behaviors. In general, the written homework problems were more difficult and have more steps compared with the online homework. The written homework counted for 8% towards the cumulative score in the course. We focus on understanding the students' behavioral differences between online and written homework problems.

As shown in Figure 4.12 and Figure 4.13, we performed hierarchical clustering analysis and correlation analysis on these behavioral variables. In Figure 4.14 we show the two student clusters for their behaviors on written homework problems using K-means clustering method. Cluster A has 327 students while cluster B has 146 students. Similar as the results from online homework problems, students in cluster A on average demonstrated more productive learning behaviors and performed significantly better than students in cluster B in both cumulative grade and final exam score (t-test with $p < 0.001$).

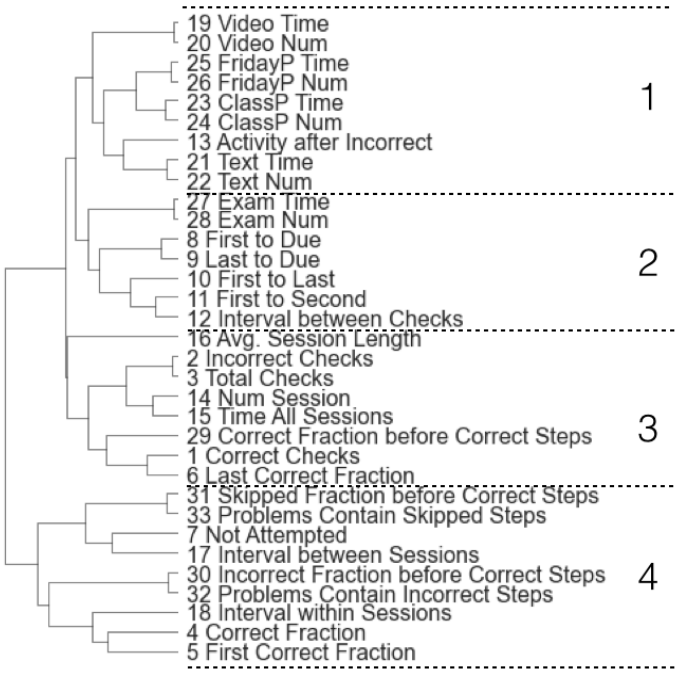
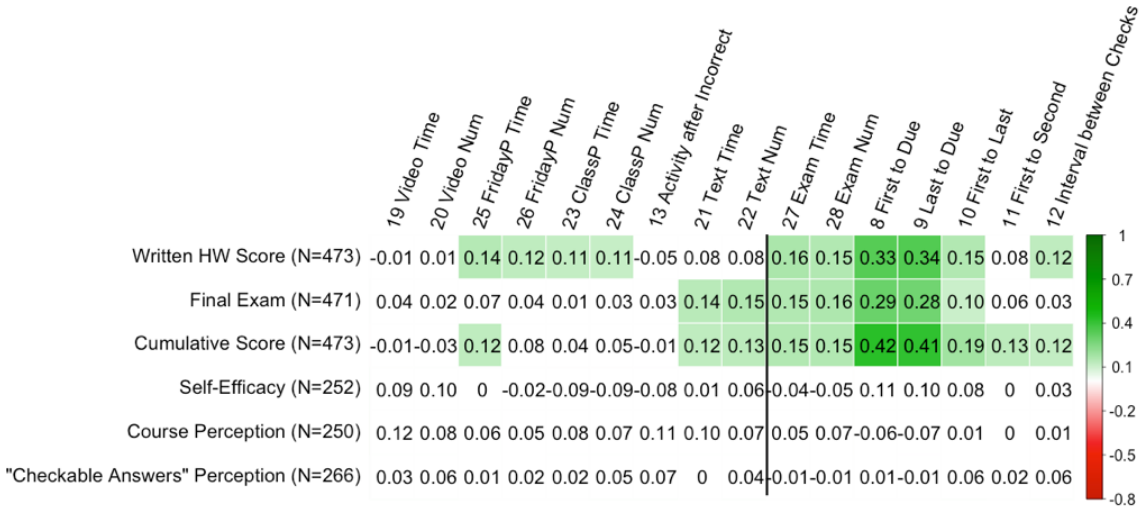
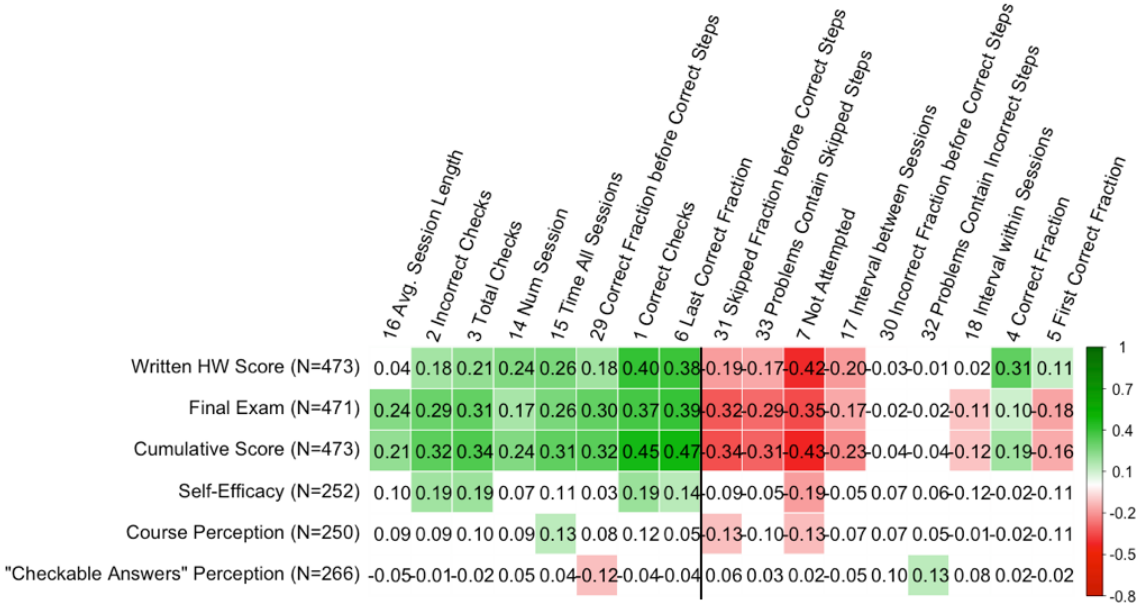


Figure 4.12 Hierarchy of the 33 behavioral variables for written homework based on their Spearman correlation distances



(a)



(b)

Figure 4.13 Spearman correlation coefficients between the 33 behavioral variables for written homework and student performance and attitudinal scores from the survey (cell background is white if p >= 0.05)

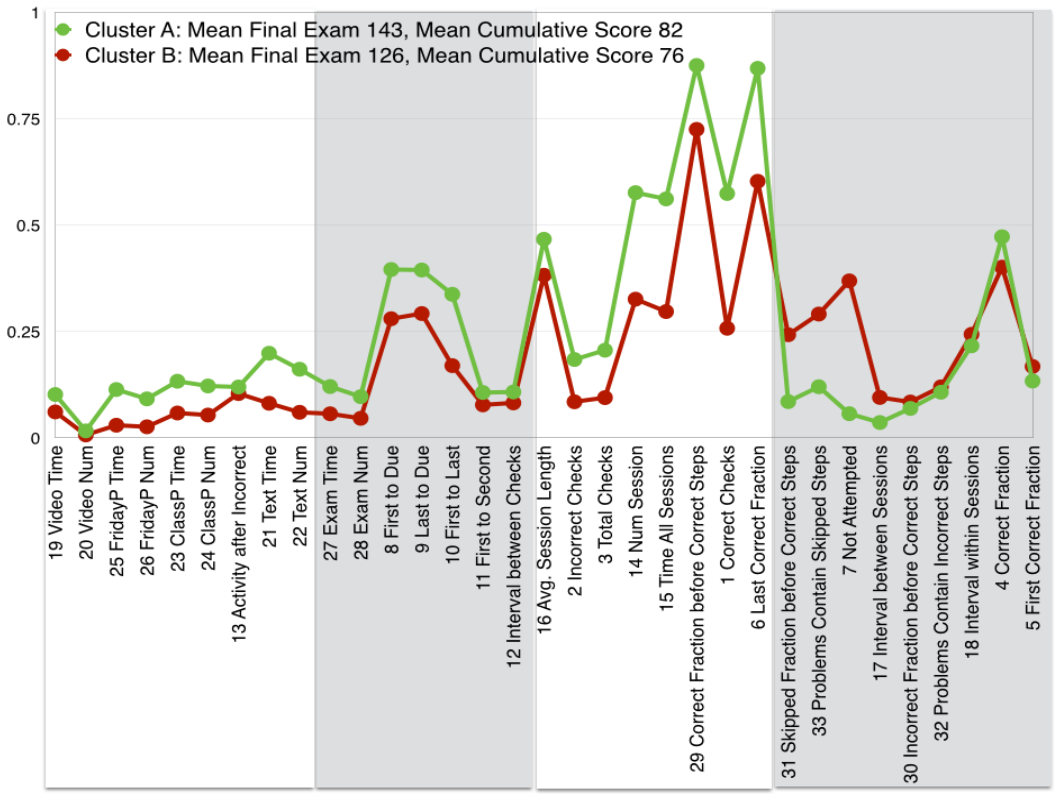
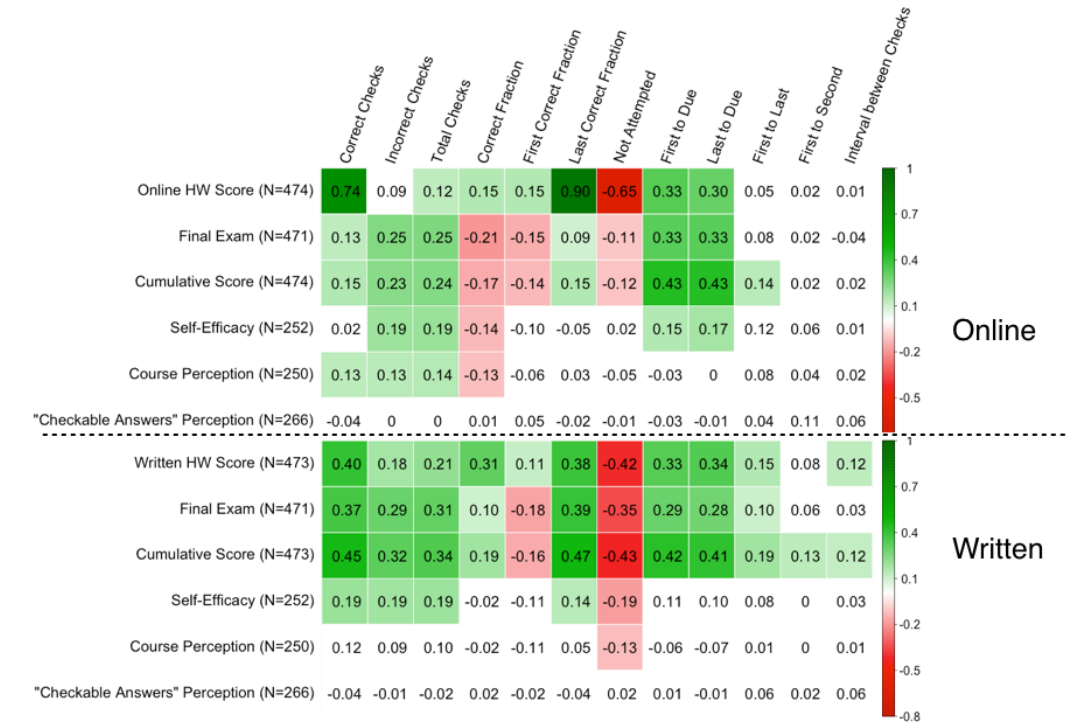


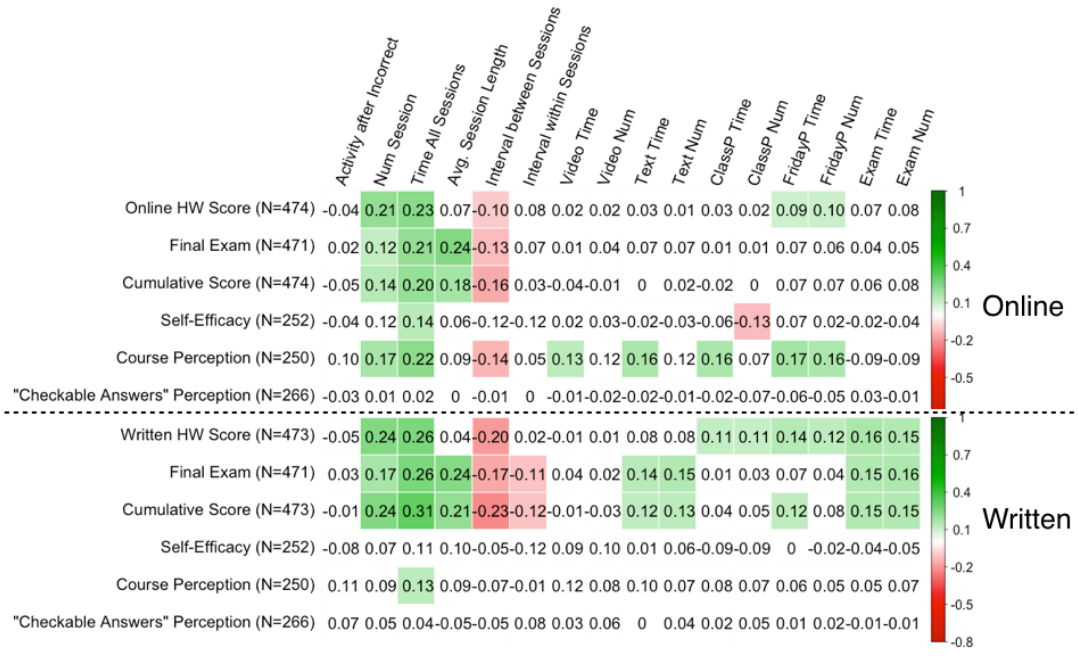
Figure 4.14 Comparison of the centroids of the two student clusters for their behaviors on written homework problems (all variables normalized to 0-1)

4.2.1 Differences in Student Behaviors for Written and Online Homework

In Figure 4.15, we aligned the correlation analysis on the same variables from online homework and written homework in order to easily discern the differences. We can see that larger number of correct checks and higher measures of checking until correct behavior are more strongly positively correlated with better performance. Also, the fraction of correct checks is positively correlated with performance in written homework, while the relationship is negative in online homework.



(a)



(b)

Figure 4.15 Comparison of student behaviors with online and written homework

The number of not attempted problems is more strongly negatively correlated with cumulative grade and final exam score, and is also negatively correlated with student self-efficacy and course perception measures.

Students' use of other resources while working on written homework including e-textbook, notes for Friday problem-solving session and in-class problems, and exam materials is more indicative of better performance compared with the use of those resources in online homework.

4.2.2 Behaviors with Problem Steps

Written homework contained more steps, and each step was designed to provide more information to the next steps. The elaborate feedback was segmented into small pieces with each piece being simple corrective feedback. We would like to understand student behaviors with these steps, so we added 5 more variables (variables 29-33) as detailed in 3.3.1. From Figure 4.16, we can see that most students follow the steps with the median of variable 29 being 0.89 (variable 29 = 1 if a student always follows the steps). Seen from Figure 4.13 (b), following the steps (variable 29) is positively correlated with student performance, which confirms our hypothesis that if students follow the steps, they would get feedback and information from previous steps, which will benefit their problem-solving in successive steps.

We had hypothesized that if the student skipped a step that probably meant they already knew how to do it. However, we found that skipping steps (variable 31 and 33) is negatively related to performance while getting incorrect steps does not matter that much (variable 30 and 32).

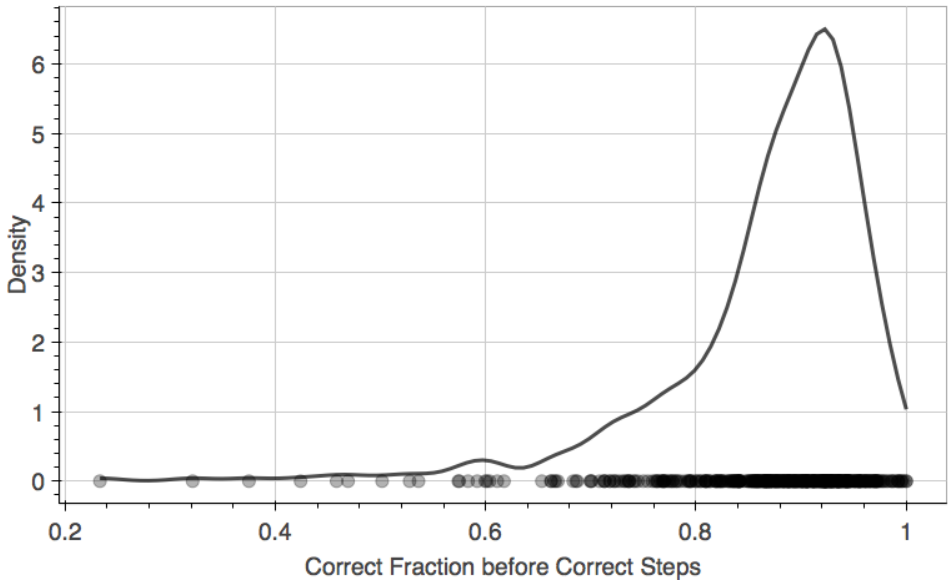


Figure 4.16 Density plot for the fraction of previous correct steps when the first time the student got any step correct (variable 29 for written homework)

CHAPTER 5. DISCUSSION

In our work, we find that higher student engagement with online immediate feedback—as measured by the numerous simple behaviors we study—is positively correlated to achievement in the course. In section 3.1.1, we had a broad hypothesis that we would identify diverse behaviors among students, and some behaviors might represent overall productive learning strategies, while others did not. Our findings indicate that productive behaviors include students checking their answers multiple times, checking the answers until correct, starting homework early, working frequently, and organizing longer and more concentrated study sessions. We identify student behavioral differences regarding the online and written homework problems, and find that following the problem steps is a behavior that is positively related to higher performance. Despite that each student could demonstrate numerous behaviors, we only identified two distinctive behavioral clusters among students. One demonstrates comparatively more productive and persistent behaviors than the other and has significantly higher performance in the course. We did not find any cluster of students who could "breeze through" the course—i.e. students who were thought to have high prior subject knowledge thus did not need to spend effort on the course while still achieved high performance in the course.

In section 3.1.1, we also had a very specific hypothesis that has a lot of grounding in the literature that answer-until-correct (AUC) behavior is positively correlated with

course performance. Our findings indeed support this hypothesis that answer-until-correct is a productive behavior that is positively related to student achievement.

We demonstrate these findings despite the surprising lack of a significant relationship between students' perceptions of the utility of the "checkable answers" and their actual behaviors engaging with this feature. Students' attitudes and self-efficacy towards the course content, on the other hand, are related to many behaviors that are strong indicators of student success in the course, which echoes findings from solely face-to-face courses.

In this discussion, we supplement our findings and make sense of them with initial qualitative analysis results from the semi-structured student interviews and the specific context of the blended learning environment.

5.1 Behaviors with Immediate Feedback

Many of the student behaviors with "checkable answers" that we measured are correlated to achievement in the course. Fewer are significantly correlated to student attitudes or perceptions. Further, some, but not all, of the relationships are in the direction we had initially expected.

The number of correct checks and the fraction of problems where students checked until correct are positively correlated with student performance. Previous literature has shown that when immediate feedback is coupled with answer-until-correct (AUC) procedure (students are forced to answer questions until they provide the correct answer), it is more effective for learning (Clariana, 1990; Persky & Pollack, 2008). In our context, though not required to always answer until correct especially in the written homework, students who did so had higher performance compared with who did not.

The number of checks, regardless of whether they are correct or incorrect, has a positive correlation with student performance. We hypothesize that this might indicate that, even if a check is incorrect, the student receives feedback on his/her response and can utilize that feedback, which may then result in higher learning outcomes. This finding is consistent the studies we have reviewed which emphasize the important role of deliberate practice in forming expertise (Ericsson, 2015; Ericsson et al., 1993). In addition, our initial qualitative analysis results confirm that students like the "checkable answers" feature because it gave them the ability to know immediately for sure whether or not they were right, and getting incorrect answers often strongly encouraged and motivated them to question and correct their initial incorrect assumptions. The positive relationship between number of checks and achievement may also be a proxy for the student's effort, which would positively contribute to performance compared with no check at all (Pintrich & de Groot, 1990). In the semi-structured interview, some students also talked about the fact that they sometimes gave up out of frustration or did not allot enough time to complete the assignment and thus had a lower number of checks. These students realized those behaviors had not been doing any good for their performance.

We find that the correlation coefficients between the fraction of problems where the student gets the answer correct on the first try and course performance are negative. This may indicate that behaviors that encompass the effort of actually checking (e.g., number of incorrect checks) and working on the problem (e.g., time between first and last check, session time) increase the students' opportunity to receive and reflect on feedback. This is more beneficial for student achievement than efficiency (getting the correct answer within the fewest possible number of checks). We had hypothesized that students

getting the correct answer on their first try might be a proxy for higher prior knowledge of the course content. Based on previous work, this would likely highly positively correlate with subsequent course performance. However, our current correlation analysis shows the opposite. Further, based on Figure 4.8, we see that those who got the correct answers on the first try tended to be those who started the homework late and submitted the homework in the last two days preceding the due date. Therefore, we recognize that getting the correct answer on the first try may represent other behaviors than having high prior knowledge, such as copying answers from peers. Although these students got most of their online homework problems correct and thus had high online homework scores, their final exam and cumulative grades appear to have been hurt by this behavior.

Variable 4 –*fraction of correct checks* is a complex and interesting one. We find that it is negatively correlated with performance for online homework problems, whereas the relationship is positive for written homework. This is because this variable contains two components each at play for the online and written homework respectively. High fraction of correct checks can occur when either the students often get the correct answer on their first try or the students often persist until correct. We now know that the former is a counter-productive behavior while the later is a productive behavior. For the written homework, checking online is optional, there was no need for the students to copy answers from others or cheat online in any other forms. Therefore, if the student has relatively large fraction of correct checks, this should mean they usually work until correct. For the same reason, the number of correct checks (variable 1) and last correct fraction (variable 6) are more strongly indicative of better performance for written homework compared with online homework.

Number of unattempted problems and skipped steps are negatively correlated to performance. This implies that students who skipped the problems or steps may lose the opportunity to learn from the feedback, and also these students were probably those who lacked proper study strategies or those who lacked the motivation to engage with the homework. This negative relationship is stronger for written homework. Our initial qualitative results also indicate that students in general thought written homework held more accountability. The fact that the written homework requires the students to submit physical copies made it easier to remember, and effort spent on written homework more strongly contributed to course performance.

5.2 Organization of Time

A set of our findings reflects student time management strategies. The time between the first check and the homework due time and the time between the last check and the homework due time are strongly correlated to students' course performance. This indicates that "early starters", on average, perform well. These students are also those who submitted the homework early compared with those who started and submitted at the last minute. This may be due to their higher level of metacognition or more mature time organization strategies. Or, this may be due to the increased amount of time necessarily allowed for feedback, reflection, and integration.

We also find that if students have longer, more frequent, and more concentrated study sessions, they tend to perform well in the course. These measures again are proxies for students' effort and engagement with the feedback. These findings corroborate with previous studies that students' level of engagement with an online course environment

strongly mediates their performance (Wei, Peng, & Chou, 2015; Zheng & Warschauer, 2015).

The interval between sessions is negatively correlated with student performance. This indicates that if students engage with the course platform more frequently, it is beneficial to their learning. We recognize that this might be due to the fact that if students waited for too long to resume their study sessions, they might have forgotten the feedback they obtained from previous sessions, thereby losing the opportunity to benefit from their reflection.

Our initial qualitative analysis results also indicate that procrastination and time management are a major factor influencing student performance. Conscientious students plan ahead—they schedule adequate time to complete assignments and generally adhere to the set schedule. Some students talked about not having enough time to do assignments because they put it off until it was too late; at least one student talked about changing their study strategy after they noticed that procrastinating was damaging their ability to get things done on time. This finding corroborates previous studies which find that procrastination and lack of focus negatively influence course performance (Streveler et al., 2003).

5.3 Self-efficacy and Perceptions

We find that some student behaviors are related to three key attitudinal scales: students' self-efficacy to perform in the course, their perceptions of the utility of the course for their future, and their perceptions of the utility of the "checkable answers". These key metrics may be related both to student performance directly (given what we know about student perceptions of instrumentality and course performance) and also to

the ways in which students use the feedback feature, which could in turn mediate their learning.

Remarkably, students' perception of the utility of the "checkable answers" feature is not actually correlated with any of their actual behaviors with that feature or their course performance. This contradicts a previous study stating that both students' perception and actual use of the interactive functions are strongly correlated to performance (Wei et al., 2015). Our finding reflects the fact that, regardless of how students perceive the benefits of the tool, it is their realized behaviors and the ways in which they engage with the feedback system that ultimately impact their performance. One possible reason that students' perceptions and their actual use of the "checkable answers" differ in this blended learning context could be peer influence and peer pressure. A student who did not perceive the answer checker as useful might still used it often because their peers used it in collaborative work. Much future work can be done to gain deeper understanding of various possible reasons. From another angle, this finding also implies that instructors and designers may need to reconsider the way they gather student perception measures.

Students' self-efficacy and their perceived utility of the course are positively correlated with many behavioral measures that approximate students' effort and engagement. The variables that are negatively correlated with student performance are also negatively correlated with self-efficacy and course perception. This shows that student attitudes operate with the same directionality as their engagement with the "checkable answers" feature in the course.

Interestingly, for online homework problems, students' perceived utility of the course to their future is positively correlated with many measures capturing the usage of other online resources. Although, in general, high usage of other online resources is not a strong indicator for engagement in this blended learning course, students who perceive this course as useful for their future still use more online resources. This may indicate that students who perceive that the course is useful for their future may also tend to perceive the online resources as more useful and therefore opt to spend more effort on it, even for the relatively easy online homework problems.

5.4 Navigating the Blended Learning Environment

In this blended learning context, we did not find much evidence for "gaming" behaviors. One reason might be that the learning environment was quite open, and students could access any available resources online and offline during the problem-solving process. In addition, correctness checking is optional for written homework. So there was no need to game the system. The students would simply give up and disengage, or cheat in other ways such as copying answers from peers rather than trick the online learning system.

We also did not find engagement with the online resources during problem-solving to be strongly correlated with performance. This may indicate that use of other online resources in the course platform is not a good measure for engagement, because in this blended learning context, students have many other offline resources such as peer discussion, tutoring sessions, and in-class problem-solving sessions to obtain help. This reflects one of the major differences between blended learning courses and entirely online courses. Our initial qualitative analysis results also confirm that students have a general

preference for face-to-face resources rather than the online resources. Many students also usually work on the homework problems in teams with other peers.

In the comparison between student behaviors for the online and written homework, we did find that because the entire assessment process for the online homework is unproctored, the assessment results could deviate from the true measure of students' actual ability on the problems. For example, more of the correct checks for the online homework came from getting the first checks correct rather than persisting until correct. We know the former is a counter-productive behavior, which could come from directly copying answers from peers. This corresponds to some instructors' concern that unproctored exams or homework assignments instrumented through the online platform may be biased or otherwise invalid (Ardid et al., 2015). Our finding clearly indicates that though this counter-productive behavior would help students get high online homework scores, their cumulative grades and final exam scores in this course would be hurt by this behavior. This finding can help instructors to provide students recommendations on proper learning behaviors that would eventually benefit their long-term learning rather than only achieving short term goals.

CHAPTER 6. IMPLICATIONS AND LIMITATIONS

In this dissertation, we tap into the poorly understood area of student behaviors using immediate feedback features on an online course platform in a blended learning environment. Our study is grounded in learning theories and relevant empirical studies on the feedback effect and student behaviors with computer-mediated feedback. Utilizing server log data that track students' every interaction with the course platform, combined with a set of students' background factors, we present in-depth analyses of student behaviors. In order to mine patterns of student behaviors and understand how these behaviors are associated with student course performance, we combine inductive and deductive methods from the traditional statistics and the machine learning communities, which are arguably two separate quantitative data analysis paradigms that are rarely presented in one study. The quantitative results are supplemented by qualitative data including semi-structured interviews and observation videos of students solving homework problems.

Understanding of the student behaviors can help instructors to provide students timely and personalized interventions and recommendations on appropriate study strategies. The fact that these recommendations are drawn from the records of students' data can make them more relevant and convincing to students. The following recommendations to students can be drawn from our findings:

1. The students should utilize the "checkable answers" to check multiple times and to reflect on the feedback.
2. The students should not focus on getting the correct answers in fewest possible checks (efficiency). Though the online homework is completely unproctored, convincing evidence from the data can be shown to students that focusing on efficiency in order to get high online homework score is negatively correlated with the final course performance. Instrumental behaviors are not beneficial for their long term learning goals. Rather, the students should be encouraged to reflect on the incorrect answers and persist to get the correct answers eventually.
3. The students should plan ahead to start the homework early, therefore allow themselves enough time to engage with the problems and reflect on the feedback.
4. The students should organize frequent and concentrated study time. Waiting for too long to resume a study session is not beneficial for learning.

Although the subject studied here is physics, it is a required course for all first-year students who could be majoring in a number of STEM fields afterwards, so the results could inform the pedagogical design of other first-year STEM classes that utilize a blended learning format.

There are a number of limitations of this study. We acknowledge that student experiences in a blended learning environment are complex. Student behaviors and performance are influenced by numerous of factors. Despite of our effort to include as many background factors as we could, there are still some factors that are only reflected in the qualitative data or not captured in any of our datasets at all, such as students' metacognitive skills. For the ones we have included in our quantitative model, some of

them are only proxy measures. For the self-reported variables measured using a survey instrument, we got a response rate of 55.88%, which means we only have data from a little over half of the students who were enrolled in this course. Our analyses have shown that students who completed the survey perform significantly better in terms of their cumulative grades and final exam scores. This result indicates that students who completed the survey and who did not might have systematic differences, and some of these differences may be unobservable factors that are not represented in our data. In other words, findings from students who responded to the survey may neither be telling the whole story, nor generalizable to the whole population of mainstream first-year students. Another limitation in terms of the generalizability of this research is that although we study the mainstream first-year students who enrolled in the regular version of the introductory physics course, and that we have a large number of students (N=474), the students were all from a top-ranking elite private research university. This may reduce the generalizability of the results to first-year students enrolled in STEM courses at all types of institutions.

Another limitation involves the assessments. We have focused on student behaviors with "checkable answers" while working on the online and written homework. These two formative assessments count for 10% of the cumulative grade. We aimed to understand how much variability of the students' final learning outcomes could be explained by these behaviors. We used the cumulative grade and final exam score as the outcome variables to represent learning outcomes. Although using final exam score and cumulative grade as outcome variables is a common practice for studying learning outcomes, we have to acknowledge that whether these summative assessments are true

measures of student learning is still a very complicated question that needs much future work. The relationship between the content of the homework problems and the exams also imposed much ambiguity.

Finally, despite of the proliferation of automatic feedback features in various technology-enabled learning systems and that the focus of this dissertation is on automatic immediate feedback, a critique to automatic online feedback mechanism argues that in-person feedback provided by faculty continues to be the best source of feedback. Instructors play a multidimensional role as course designer and organizer, discussion facilitator, social support, and technology facilitator (Hung & Chou, 2015). They provide in-person feedback and support, and facilitate student-student and student-instructor communication which are strongly related to higher learner engagement, motivation, and satisfaction (Dixson, 2012; Imlawi, Gregg, & Karimi, 2015; Stephens & Clement, 2015). On the contrary, some other researchers argue that many students prefer to request feedback from a computer rather than from a person due to self-esteem and public and private self-consciousness (Kluger & Adler, 1993). Along this line, some assessment systems can now combine manual feedback from the instructors and TAs with automatic feedback to take the advantages of both (Ihantola, Ahoniemi, Karavirta, & Seppälä, 2010). To some extent, blended learning is already a good solution to this critique in that it encapsulates both automatic online feedback and face-to-face feedback from instructors, TAs, and peer students—and we see students extensively using both. In addition, by analyzing and understanding student behaviors with automatic immediate feedback, we provide instructors an anchor for them to provide students feedback, interventions, and recommendations on productive learning strategies. This is to say that

the understanding of students' usage of automatic feedback in turn serves to improve the quality of instructors' in-person feedback.

The next chapter concludes this study.

CHAPTER 7. CONCLUSION

This study utilizes rich quantitative data to address the need to understand student use of computerized immediate feedback in a blended learning setting. Using a mix of inductive and deductive methods, we generated a variety of variables to describe students' behaviors with "checkable answers" and their use of online resources while working on the online and written homework problems. We proposed a set of recommendations that instructors can use to help students adopt productive study strategies. Many of our findings are corroborated by previous studies in either face-to-face or online learning environments. In the future, when our quantitative data analysis methodology is adopted in real time, instructors can monitor student behaviors and provide timely and personalized feedback, and students can also see timely evidence of their own behaviors and aggregated behaviors from their peers, which is more relevant and more convincing.

There are several aspects where we can lay out our plans for future work. For the quantitative analysis, we plan to test several other quantitative data analysis techniques. For example, using density-based clustering methods such as DBSCAN can potentially help us identify student behavioral clusters from a different perspective. We will also continue the qualitative analysis of the semi-structured interview and observation videos of student problem-solving process to better understand why students demonstrate certain behaviors. Several findings lend themselves to further development. For example,

although in this blended learning context, we did not find intensive use of online resources such as e-textbook and problem-solving videos strongly relating to course performance, our descriptive statistics in Appendix E reveal that students spent more time on e-textbook than videos. Further research can be done to understand the affordance of the different formats of online resources therefore to improve the design of useful learning materials.

BIBLIOGRAPHY

BIBLIOGRAPHY

Aleven, V., & Koedinger, K. R. (2002). Limitations of student control: Do students know when they need help? In *Intelligent Tutoring Systems* (Vol. 1839, pp. 292–303). Springer Berlin Heidelberg.

Aragon, S. R., Johnson, S. D., & Shaik, N. (2002). The influence of learning style preferences on student success in online versus face-to-face environments. *The American Journal of Distance Education*, *16*(4), 227–243.

Ardid, M., Gómez-Tejedor, J. A., Meseguer-Dueñas, J. M., Riera, J., & Vidaurre, A. (2015). Online exams for blended assessment. Study of different application methodologies. *Computers & Education*, *81*, 296–303. http://doi.org/10.1016/j.compedu.2014.10.010

Babyak, M. A. (2004). What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine*, *66*(3), 411–421.

Baker, R. S., Corbett, A. T., & Koedinger, K. R. (2004). Detecting student misuse of intelligent tutoring systems. In *Intelligent tutoring systems* (Vol. 3220, pp. 531–540). Springer Berlin Heidelberg.

Baker, R. S., Corbett, A. T., Koedinger, K. R., & Wagner, A. Z. (2004). Off-task behavior in the cognitive tutor classroom: when students game the system. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 383–390). ACM.

Balzer, W. K., Doherty, M. E., & O'Connor, R. (1989). Effects of cognitive feedback on performance. *Psychological Bulletin*, *106*(3), 410.

Beichner, R. J., & Saul, J. M. (2003). Introduction to the SCALE-UP (student-centered activities for large enrollment undergraduate programs) project. In *Proceedings of the International School of Physics "Enrico Fermi."* Varenna, Italy.

Beichner, R. J., Saul, J. M., Abbott, D. S., Morse, J. J., Deardorff, D., Allain, R. J., … Risley, J. S. (2007). The student-centered activities for large enrollment undergraduate programs (SCALE-UP) project. *Research-Based Reform of University Physics*, *1*(1), 2–39.

Bele, J., & Rugelj, J. (2007). Blended learning - an opportunity to take the best of both worlds. *International Journal of Emerging Technologies in Learning (iJET)*, *2*(3).

Bernard, R. M., Borokhovski, E., Schmid, R. F., Tamim, R. M., & Abrami, P. C. (2014). A meta-analysis of blended learning and technology use in higher education: from the general to the applied. *Journal of Computing in Higher Education*, *26*(1), 87–122.

Bishop, C. M. (2007). *Pattern Recognition and Machine Learning*. Springer.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7–74. http://doi.org/10.1080/0969595980050102

Blumner, H. N., & Richards, H. C. (1997). Study habits and academic achievement of engineering students. *Journal of Engineering Education*, *86*(2), 125–132.

Bowen, W. G., Chingos, M. M., Lack, K. A., & Nygren, T. I. (2014). Interactive learning online at public universities: Evidence from a six-campus randomized trial. *Journal of Policy Analysis and Management*, *33*(1), 94–111.

Bower, M., Dalgarno, B., Kennedy, G. E., Lee, M. J. W., & Kenney, J. (2015). Design and implementation factors in blended synchronous learning environments: Outcomes from a cross-case analysis. *Computers & Education*, *86*, 1–17. http://doi.org/10.1016/j.compedu.2015.03.006

Brackbill, Y., Adams, G., & Reaney, T. P. (1967). A parametric study of the delay-retention effect. *Psychological Reports*, *20*(2), 433–434.

Brackbill, Y., Bravos, A., & Starr, R. H. (1962). Delay-improved retention of a difficult task. *Journal of Comparative and Physiological Psychology*, *55*(6), 947.

Brackbill, Y., & Kappy, M. S. (1962). Delay of reinforcement and retention. *Journal of Comparative and Physiological Psychology*, *55*(1), 14.

Bransford, J., Brown, A. L., & Cocking, R. R. (Eds.). (1999). *How people learn: Brain, mind, experience, and school*. Washington, D.C.: National Academies Press.

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3), 199–231. http://doi.org/10.1214/ss/1009213726

Breslow, L., Pritchard, D. E., DeBoer, J., Stump, G. S., Ho, A. D., & Seaton, D. T. (2013). Studying learning in the worldwide classroom: Research into edX's first MOOC. *Research & Practice in Assessment*, *8*, 13–25.

Butler, A., Karpicke, J., & Roediger III, H. (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology: Applied*, *13*(4), 273.

Butler, R. (1988). Enhancing and undermining intrinsic motivation: The effects of task-involving and ego-involving evaluation on interest and performance. *British Journal of Educational Psychology*, *58*(1), 1–14.

Chase, J. A., & Houmanfar, R. (2009). The differential effects of elaborate feedback and basic feedback on student performance in a modified, personalized system of instruction course. *Journal of Behavioral Education*, *18*(3), 245–265. http://doi.org/10.1007/s10864-009-9089-2

Chen, X., Vorvoreanu, M., & Madhavan, K. P. (2014). Mining social media data for understanding students' learning experiences. *IEEE Transactions on Learning Technologies*, *7*(3), 246–259. http://doi.org/10.1109/TLT.2013.2296520

Clariana, R. B. (1990). A comparison of answer until correct feedback and knowledge of correct response feedback under two conditions of contextualization. *Journal of Computer-Based Instruction*.

Clariana, R. B. (1999). Differential memory effects for immediate and delayed feedback: A delta rule explanation of feedback timing effects.

Clariana, R. B., Wagner, D., & Murphy, L. C. R. (2000). Applying a connectionist description of feedback timing. *Educational Technology Research and Development*, *48*(3), 5–22.

Corbett, A. T., & Anderson, J. R. (2001). Locus of feedback control in computer-based tutoring: Impact on learning rate, achievement and attitudes. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 245–252). ACM.

Coutinho, S. (2008). Self-efficacy, metacognition, and performance. *North American Journal of Psychology*, *10*(1), 165.

DeBoer, J., Ho, A. D., Stump, G. S., & Breslow, L. (2014). Changing "course" reconceptualizing educational variables for massive open online courses. *Educational Researcher*, *43*(2), 74–84. http://doi.org/10.3102/0013189X14523038

DeBoer, J., Stump, G. S., Seaton, D., & Breslow, L. (2013). Diversity in MOOC students' backgrounds and behaviors in relationship to performance in 6.002 x. In *Proceedings of the Sixth Learning International Networks Consortium Conference*.

Dihoff, R. E., Brosvic, G. M., & Epstein, M. L. (2003). The role of feedback during academic testing: The delay retention effect revisited. *The Psychological Record*, *53*(4).

Dihoff, R. E., Brosvic, G. M., Epstein, M. L., & Cook, M. J. (2004). Provision of feedback during preparation for academic testing: Learning is enhanced by immediate but not delayed feedback. *Psychological Record*, *54*(2), 207–232.

Dixson, M. D. (2012). Creating effective student engagement in online courses: What do students find engaging? *Journal of the Scholarship of Teaching and Learning*, *10*(2), 1–13.

Earley, P. C., Northcraft, G. B., Lee, C., & Lituchy, T. R. (1990). Impact of process and outcome feedback on the relation of goal setting to task performance. *Academy of Management Journal*, *33*(1), 87–105.

Elder, B. L., & Brooks, D. W. (2008). Simple versus elaborate feedback in a nursing science course. *Journal of Science Education and Technology*, *17*(4), 334–340. http://doi.org/10.1007/s10956-008-9103-9

Epstein, M. L., & Brosvic, G. M. (2002). Students prefer the immediate feedback assessment technique. *Psychological Reports*, *90*(3c), 1136–1138.

Epstein, M. L., Lazarus, A., Calvano, T., Matthews, K., Hendel, R., Epstein, B., & Brosvic, G. (2010). Immediate feedback assessment technique promotes learning and corrects inaccurate first responses. *The Psychological Record*, *52*(2).

Ericsson, K. A. (2004). Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. *Academic Medicine*, *79*(10), S70–S81.

Ericsson, K. A. (2015). Acquisition and maintenance of medical expertise: A perspective from the expert-performance approach with deliberate practice. *Academic Medicine*, *90*(11), 1471–1486.

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, *100*(3), 363.

Garrison, D. R., & Kanuka, H. (2004). Blended learning: Uncovering its transformative potential in higher education. *The Internet and Higher Education*, *7*(2), 95–105. http://doi.org/10.1016/j.iheduc.2004.02.001

Google. (2015). Google analytics cookie usage on websites. Retrieved October 8, 2015, from https://developers.google.com/analytics/devguides/collection/analyticsjs/cookie-usage

Gordijn, J., & Nijhof, W. J. (2002). Effects of complex feedback on computer-assisted modular instruction. *Computers & Education*, *39*(2), 183–200.

Halawa, S., Greene, D., & Mitchell, J. (2014). Dropout prediction in MOOCs using learner activity features. *Proceedings of the European MOOC Stakeholder Summit (EMOOCS 2014), Lausanne, Switzerland*.

Halder, S., Saha, S., & Das, S. (2015). Computer based self-pacing instructional design approach in learning with respect to gender as a variable. In *Information Systems Design and Intelligent Applications* (Vol. 340, pp. 37–47). Springer India.

Hall, R. M., & Sandler, B. R. (1982). *The classroom climate: A chilly one for women?*

Harackiewicz, J. M., Manderlink, G., & Sansone, C. (1984). Rewarding pinball wizardry: Effects of evaluation and cue value on intrinsic interest. *Journal of Personality and Social Psychology*, *47*(2), 287.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *28*(1), 100–108. http://doi.org/10.2307/2346830

Hastie, T., Tibshirani, R., Friedman, J., & Franklin, J. (2013). *The elements of statistical learning: data mining, inference and prediction* (2nd ed.). Springer.

Hattie, J. (2013). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. http://doi.org/10.3102/003465430298487

Hung, M.-L., & Chou, C. (2015). Students' perceptions of instructors' roles in blended and online learning environments: A comparative study. *Computers & Education*, *81*, 315–325. http://doi.org/10.1016/j.compedu.2014.10.022

Husman, J., Derryberry, W. P., Crowson, H. M., & Lomax, R. (2004). Instrumentality, task value, and intrinsic motivation: Making sense of their independent interdependence. *Contemporary Educational Psychology*, *29*(1), 63–76.

Ihantola, P., Ahoniemi, T., Karavirta, V., & Seppälä, O. (2010). Review of recent systems for automatic assessment of programming assignments. In *Proceedings of the 10th Koli Calling International Conference on Computing Education Research* (pp. 86–93). New York, NY, USA: ACM. http://doi.org/10.1145/1930464.1930480

Imlawi, J., Gregg, D., & Karimi, J. (2015). Student engagement in course-based social networks: The impact of instructor credibility and use of communication. *Computers & Education*, *88*, 84–96. http://doi.org/10.1016/j.compedu.2015.04.015

Interactive Advertising Bureau. (2009). IAB - guidelines, specifications & best practices. Retrieved October 8, 2015, from http://www.iab.net/guidelines

Kizilcec, R. F., Piech, C., & Schneider, E. (2013). Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In *Proceedings of the third international conference on learning analytics and knowledge* (pp. 170–179). ACM.

Kluger, A. N., & Adler, S. (1993). Person- versus computer-mediated feedback. *Computers in Human Behavior*, *9*(1), 1–16. http://doi.org/10.1016/0747-5632(93)90017-M

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*(2), 254.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*(6), 1121.

Krumm, A. E., D'Angelo, C., Podkul, T. E., Feng, M., Yamada, H., Beattie, R., … Thorn, C. (2015). Practical measures of learning behaviors. In *Proceedings of the Second (2015) ACM Conference on Learning@ Scale* (pp. 327–330). ACM.

Kuhn, D. (2000). Metacognitive development. *Current Directions in Psychological Science*, *9*(5), 178–181. http://doi.org/10.1111/1467-8721.00088

Kukla, A. (2000). *Social constructivism and the philosophy of science*. Psychology Press.

Kulhavy, R. W. (1977). Feedback in written instruction. *Review of Educational Research*, *47*(2), 211–232. http://doi.org/10.2307/1170128

Kulhavy, R. W., & Anderson, R. C. (1972). Delay-retention effect with multiple-choice tests. *Journal of Educational Psychology*, *63*(5), 505.

Kulhavy, R. W., & Stock, W. A. (1989). Feedback in written instruction: The place of response certitude. *Educational Psychology Review*, *1*(4), 279–308.

Kulhavy, R. W., White, M. T., Topp, B. W., Chan, A. L., & Adams, J. (1985). Feedback complexity and corrective efficiency. *Contemporary Educational Psychology*, *10*(3), 285–291.

Kulik, J. A., & Kulik, C.-L. C. (1988). Timing of feedback and verbal learning. *Review of Educational Research*, *58*(1), 79–97.

Lack, K. A. (2013). *Current status of research on online learning in postsecondary education*. ITHAKA S+R.

Locke, E. A., & Latham, G. P. (1990). *A theory of goal setting & task performance.* Prentice-Hall, Inc.

Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American Psychologist*, *57*(9), 705.

Mathan, S. A., & Koedinger, K. R. (2002). An empirical assessment of comprehension fostering features in an intelligent tutoring system. In *Intelligent Tutoring Systems* (Vol. 2363, pp. 330–343). Springer Berlin Heidelberg.

Means, B., Bakia, M., & Murphy, R. (2014). *Learning online: What research tells us about whether, when and how* (1st ed.). Routledge.

Means, B., Toyama, Y., Murphy, R., Bakia, M., & Jones, K. (2009). Evaluation of evidence-based practices in online learning: A meta-analysis and review of online learning studies. *US Department of Education*.

Morris, L. V., Finnegan, C., & Wu, S. (2005). Tracking student behavior, persistence, and achievement in online courses. *The Internet and Higher Education*, *8*(3), 221–231. http://doi.org/10.1016/j.iheduc.2005.06.009

Narciss, S., & Huth, K. (2004). How to design informative tutoring feedback for multimedia learning. *Instructional Design for Multimedia Learning*, 181–195.

Newman, R. S. (1994). Adaptive help seeking: A strategy of self-regulated learning. In *Self-regulation of learning and performance: Issues and educational applications* (pp. 283–301). Lawrence Erlbaum Associates, Inc.

Pellegrino, J. W., Chudowsky, N., Glaser, R., & National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, D.C.: National Academy Press, Board on Testing and Assessment, Center for Education, Division of Behavioral and Social Sciences and Education.

Persky, A. M., & Pollack, G. M. (2008). Using answer-until-correct examinations to provide immediate feedback to students in a pharmacokinetics course. *American Journal of Pharmaceutical Education*, *72*(4).

Phye, G. D., & Bender, T. (1989). Feedback complexity and practice: Response pattern analysis in retention and transfer. *Contemporary Educational Psychology*, *14*(2), 97–110.

Pintrich, P. R., & de Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology*, *82*(1), 33–40. http://doi.org/10.1037/0022-0663.82.1.33

Pintrich, P. R., Smith, D. A. F., Garcia, T., & Mckeachie, W. J. (1993). Reliability and predictive validity of the motivated strategies for learning questionnaire (Mslq). *Educational and Psychological Measurement*, *53*(3), 801–813. http://doi.org/10.1177/0013164493053003024

Powell, R., Conway, C., & Ross, L. (1990). Effects of student predisposing characteristics on student success. *International Journal of E-Learning & Distance Education*, *5*(1), 5–19.

Richards, L. G. (2001). Further studies of study skills and study habits. In *Frontiers in Education Conference (FIE)*.

Riley, D. M., & Pawley, A. L. (2011). Complicating difference: Exploring and exploding three myths of gender and race in engineering education. In *American Society for Engineering Education Annual Conference*.

Rokach, L. (2010). A survey of clustering algorithms. In *Data Mining and Knowledge Discovery Handbook* (pp. 269–298). Springer.

Roll, I., Aleven, V., McLaren, B. M., & Koedinger, K. R. (2011). Improving students' help-seeking skills using metacognitive feedback in an intelligent tutoring system. *Learning and Instruction*, *21*(2), 267–280.

Roll, I., Baker, R. S., Aleven, V., & Koedinger, K. R. (2014). On the benefits of seeking (and avoiding) help in online problem-solving environments. *Journal of the Learning Sciences*, *23*(4).

Ryan, A. M., Patrick, H., & Shim, S. (2005). Differential profiles of students identified by their teacher as having avoidant, appropriate, or dependent help-seeking tendencies in the classroom. *Journal of Educational Psychology*, *97*(2), 275.

Schmidt, R. A., Young, D. E., Swinnen, S., & Shapiro, D. C. (1989). Summary knowledge of results for skill acquisition: Support for the guidance hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(2), 352.

Schroth, M. L. (1992). The effects of delay of feedback on a delayed concept formation transfer task. *Contemporary Educational Psychology*, *17*(1), 78–82.

Schwartz, D. L., & Arena, D. (2013). *Measuring what matters most: Choice-based assessments for the digital age*. MIT Press.

Seaton, D. T., Bergner, Y., Chuang, I., Mitros, P., & Pritchard, D. E. (2014). Who does what in a massive open online course? *Communications of the ACM*, *57*(4), 58–65. http://doi.org/10.1145/2500876

Seaton, D. T., Bergner, Y., & Pritchard, D. E. (2013). Exploring the relationship between course structure and etext usage in blended and open online courses. In *Proceedings of the 6th International Conference on Educational Data Mining*.

Seaton, D. T., Nesterko, S., Mullaney, T., Reich, J., & Ho, A. (2014). In-depth characterizing video use in the catalogue of MITx MOOCs. *Elearning Papers*.

Sebatane, E. M. (1998). Assessment and classroom learning: A response to Black & Wiliam. *Assessment in Education*, *5*(1), 123–130.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*(1), 153–189. http://doi.org/10.3102/0034654307313795

Silva, E., & White, T. (2013). Pathways to improvement: Using psychological strategies to help college students master developmental math. *Carnegie Foundation for the Advancement of Teaching*.

Stephens, A. L., & Clement, J. J. (2015). Use of physics simulations in whole class and small group settings: Comparative case studies. *Computers & Education*, *86*, 137–156. http://doi.org/10.1016/j.compedu.2015.02.014

Streveler, R. A., Hoeglund, T., & Stein, C. (2003). The study strategies of academically successful students at the Colorado school of mines. In *Frontiers in Education (FIE)*. IEEE.

Thompson, W. B. (1998). Metamemory accuracy: effects of feedback and the stability of individual differences. *The American Journal of Psychology*, *111*(1), 33–42.

Timmers, C. F., & Veldkamp, B. (2011). Attention paid to feedback provided by a computer-based assessment for learning on information literacy. *Computers & Education*, *56*(3), 923–930. http://doi.org/10.1016/j.compedu.2010.11.007

Timmers, C. F., Walraven, A., & Veldkamp, B. P. (2015). The effect of regulation feedback in a computer-based formative assessment on information problem solving. *Computers & Education*, *87*, 1–9. http://doi.org/10.1016/j.compedu.2015.03.012

Twigg, C. A. (2003). Improving learning and reducing costs: New models for online learning. *EDUCAUSE Review*, *38*(5).

Vockell, E. (2004). Educational psychology: A practical approach. *Purdue University, on-Line Book*. Retrieved from http://education.purduecal.edu/Vockell/EdPsyBook/

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*.

Vygotsky, L. S. (1986). *Thought and language*. (A. Kozulin, Ed.) (Rev'd). The MIT Press.

Wei, H., Peng, H., & Chou, C. (2015). Can more interactivity improve learning achievement in an online course? Effects of college students' perception and actual use of a course-management system on their learning achievement. *Computers & Education*, *83*, 10–21. http://doi.org/10.1016/j.compedu.2014.12.013

Weinstein, C. E. (1996). Learning how to learn: An essential skill for the 21st century. *Educational Record*, *66*(4), 49–52.

Weinstein, C. E., & Hume, L. M. (1998). *Study strategies for lifelong learning*. American Psychological Association.

Wilcox, R. R. (1982). Some empirical and theoretical results on an answer-until-correct scoring procedure. *British Journal of Mathematical and Statistical Psychology*, *35*(1), 57–70.

Wilson, J. M. (1994). The CUPLE physics studio. *The Physics Teacher*, *32*(9), 518–523.

Yang, D., Sinha, T., Adamson, D., & Rosé, C. P. (2013). Turn on, tune in, drop out: Anticipating student dropouts in massive open online courses. In *Proceedings of the 2013 NIPS Data-Driven Education Workshop*.

Zeldin, A. L., & Pajares, F. (2000). Against the odds: Self-efficacy beliefs of women in mathematical, scientific, and technological careers. *American Educational Research Journal*, *37*(1), 215–246.

Zhao, Y., & Breslow, L. (2013). *Literature review on hybrid/blended learning*. Retrieved from http://tll.mit.edu/sites/default/files/library/Blended_Learning_Lit_Reveiw.pdf

Zheng, B., & Warschauer, M. (2015). Participation, interaction, and academic achievement in an online discussion environment. *Computers & Education*, *84*, 78–89. http://doi.org/10.1016/j.compedu.2015.01.008

APPENDICES

Appendix A "Checkable Answers" Survey

Part I. Opinions about "checkable answers" on homework problems:

1. When you do PHYS101 homework problems, do you primarily do them by yourself or with other students?
    - By myself
    - With other student(s) in PHYS101
    - With other student(s) who took PHYS101 previously
    - With student(s) who tested out of PHYS101
    - Other: _____
2. On average, how long does it take you to do a PHYS101 problem set?
    - 1-2 hours
    - 3-4 hours
    - 5-6 hours
    - 6-7 hours
    - Over 7 hours
2. When do you typically start working on a PHYS101 problem set?
    - A week before it is due
    - 3-5 days before it is due
    - 1-2 days before it is due
    - The night before it is due
3. When you do PHYS101 homework problems, when do you check your answer?
    - After I do each part of the problem, I check my answer

- After I do the entire problem, I check my answers
- Not applicable

4. If you get a homework problem wrong on the first try, how many times will you keep trying to get the right answer?
    - I'll try 2 or 3 times
    - I'll try 4 or 5 times
    - I'll try 6 or more times
    - I keep working on the problem until I get it right
5. If you can't get the right answer on a homework problem, what do you do most often?
    - I go back to the reading summary or the text
    - I watch a problem-solving video
    - I ask another student in the class
    - I go to TA office hours
    - Other: _____
6. What do you like about the checkable answer feature for the homework problems? Please indicate your level of agreement or disagreement with each of the following statement on a scale of 1-7. 1 is strongly agree, 4 is neutral, and 7 is strongly disagree.
    - Helps build my confidence that I will do well in PHYS101.
    - Helps build my confidence that I will do well on the PHYS101 exams.
    - Contributes to my knowledge of the topics in PHYS101.
    - Helps me check for errors in my knowledge of topics in PHYS101.

- Helps me check for errors in my math.
- Makes learning easier.
- When I can see the answer, I get less frustrated.
- When I get the right answer, I'm sure I have learned the material.
- When I get the wrong answer, the checkable answer feature motivates me to find the right answer.
- Helps me retain what I have learned.
- Reinforces my correct understanding of the problem.
- Reduces my misconceptions about the material.
- Reduces the chance I will get this problem or one like it wrong in the future.

7. If there are other reasons why you like or do not like the checkable answer feature, please tell us:
8. How do you think the checkable answer feature could be improved? [check all that apply]:
    - I would like to get hints about what I am doing wrong.
    - I would like to be told what resources I could use. (e.g., reading summaries, videos)
    - I would like to be told why my answer is correct.
    - Other: _____

Part II: Opinions about "checkable answers" on pre-class reading questions

1. How often do you do the pre-class reading questions?
    - Every week

- Most every week
- Occasionally
- Never

2. When you do the pre-class reading questions, do you primarily do them by yourself or with other students?
    - By myself
    - With other student(s) in PHYS101
    - With other student(s) who took PHYS101 previously
    - With student(s) who tested out of PHYS101
    - Other: _____
3. What do you do if you get a pre-class reading question wrong on the first try?
    - I show the solution and then put it in.
    - I try to answer the question several more times before I show the solution.
    - I work on the question until I answer correctly.
4. What do you like about the checkable answer feature for the pre-class reading questions? Please indicate your level of agreement or disagreement with each of the following statement on a scale of 1-7. 1 is strongly agree, 4 is neutral, and 7 is strongly disagree.
    - Helps build my confidence that I will do well in PHYS101.
    - Helps build my confidence that I will do well on the PHYS101 exams.
    - Contributes to my knowledge of the topics in PHYS101.
    - Helps me check for errors in my knowledge of topics in PHYS101.
    - Helps me check for errors in my math.

- Makes learning easier.
- When I can see the answer, I get less frustrated
- When I get the right answer, I'm sure I have learned the material.

Part III: Attitudes about physics

1. The following statements concern your aspirations after PHYS101. Please indicate your level of agreement or disagreement with each of the following statement on a scale of 1-7. 1 is strongly agree, 4 is neutral, and 7 is strongly disagree.
    - I will major in physics for my undergraduate degree.
    - I will work in a job in the field of physics as my first full- time occupation after university.
    - I will work in physics at some point during my career.
    - I plan on continuing my education in physics after my undergraduate degree.
    - I will major in a field related to physics, science, engineering, technology, or math for my undergraduate degree.
    - I feel an obligation to work in physics, science, engineering, technology, math or a related field at some point after university.
    - I will work in a field related to physics, science, engineering, technology, or math at some point in my career.
    - I feel an obligation to my family to work in physics, science, engineering, technology, math or a related field.

2. The following statements convey beliefs about the relationship between PHYS101 and your future. Please indicate your level of agreement or disagreement with each of the following statement on a scale of 1-7. 1 is strongly agree, 4 is neutral, and 7 is strongly disagree.
    - I will use the information I learn in PHYS101 in other classes I will take in the future.
    - What I learn in PHYS101 will be important for my future occupational success.
    - I will not use what I learn in PHYS101.
    - The grade I get in PHYS101 will not affect my ability to continue on with my education.
    - I will use the information I learn in PHYS101 in the future.
    - The grade I get in PHYS101 will not be important for my future academic success.
    - I must pass PHYS101 in order to reach my academic goals.
    - The grade I get in PHYS101 will affect my future.
3. The following statements convey attitudes about PHYS101. Please indicate your level of agreement or disagreement with each of the following statement on a scale of 1-7. 1 is strongly agree, 4 is neutral, and 7 is strongly disagree.
    - I believe I will receive an excellent grade in PHYS101.
    - I am certain I can understand the most difficult material presented in the readings for PHYS101.
    - I am confident I can understand the basic concepts taught in PHYS101.

- I am confident I can understand the most complex material presented by the instructor in PHYS101.
- I expect to do well in PHYS101.
- I am certain I can master the skills being taught in PHYS101.
- I am confident I can do an excellent job on the assignments and tests in PHYS101.
- Considering the difficulty of this course, the teacher, and my skills, I think I will do well in PHYS101.

4. The following questions ask about your attitude towards the social component of PHYS101. Please indicate your level of agreement or disagreement with each of the following statement on a scale of 1-7.
    - The social aspect of the class is important to me.
    - I enjoy helping my fellow students.
    - I intend to share my expertise with other students in this class.
    - I would like to provide help as a tutor to students in future classes.
5. How many other "blended learning" classes have you taken, both in high school and at your current university? We define blended learning as a class that has a significant online component, but in which the instructor and students still meet face-to-face.
    - 0
    - 1-3
    - 4-6
    - Over 6

Appendix B Detailed Survey Results on the Three Measures Used

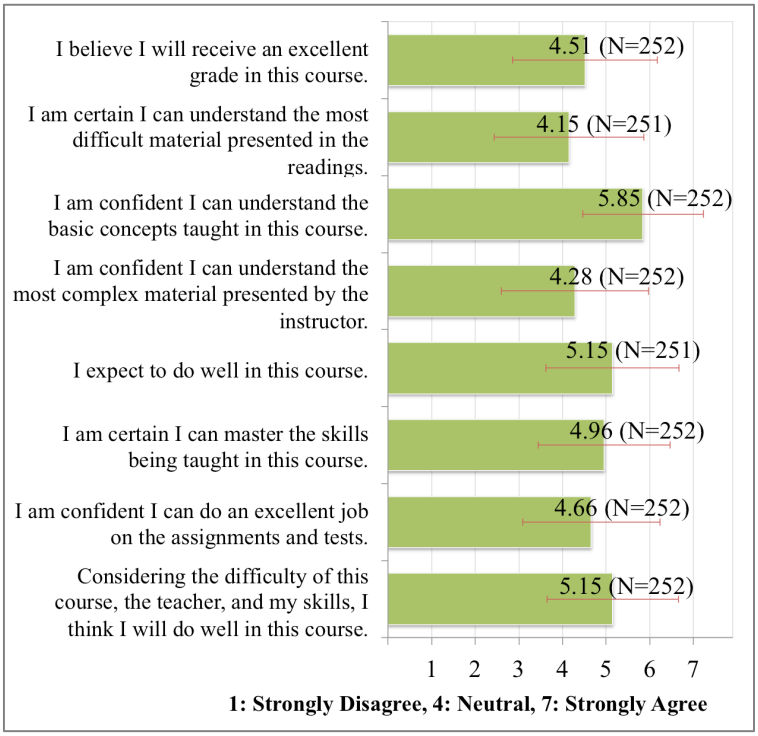


Figure B 1 Scale for students' self-efficacy to perform in the course (8 items)

Table B 1 Relationship of students' scale score on self-efficacy the their Course performance level

	Mean Scale Score of Self-Efficacy	F-value	P-value
Students in the high performance group for cumulative grade (N=64)	5.542	27.791	<0.001**
Students in the medium performance group for cumulative grade (N=152)	4.898		
Students in the low performance group for cumulative grade (N=50)	3.779		
*p <0.05, ** p < 0.01 Performance level: high [90, 100], medium [80, 90), low [0, 80)			

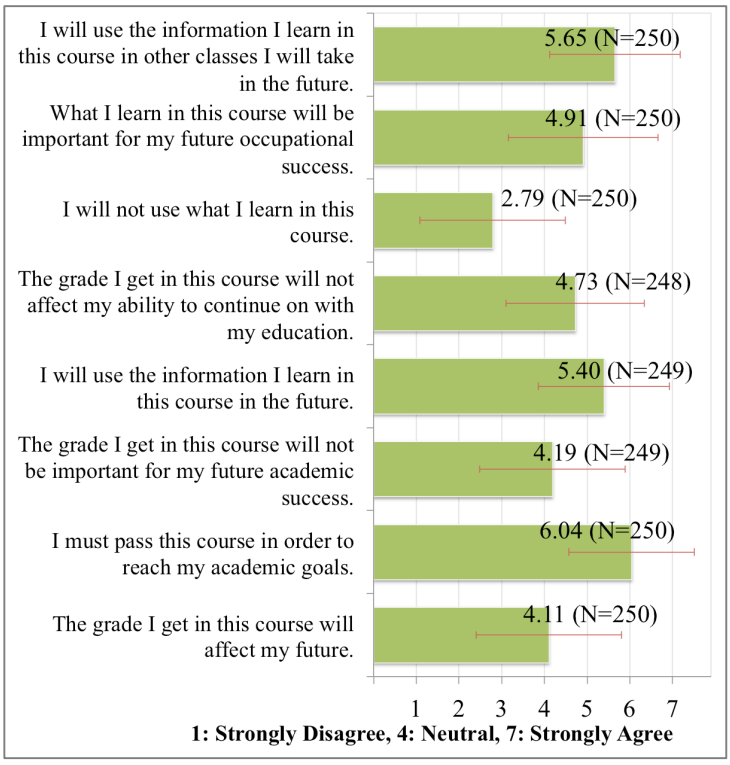


Figure B 2 Scale for students' perception of the utility of the PHYS101 course for their future (8 items)

Table B 2 Relationship of students' scale score on course perception and their course performance level

	Mean Scale Score of Perceived Utilities	**F-value**	**P-value**
Students in the high performance group for cumulative grade (N=64)	5.217	5.044	0.007**
Students in the medium performance group for cumulative grade (N=152)	5.383		
Students in the low performance group for cumulative grade (N=50)	4.754		
*p <0.05, ** p < 0.01 Performance level: high [90, 100], medium [80, 90), low [0, 80)			

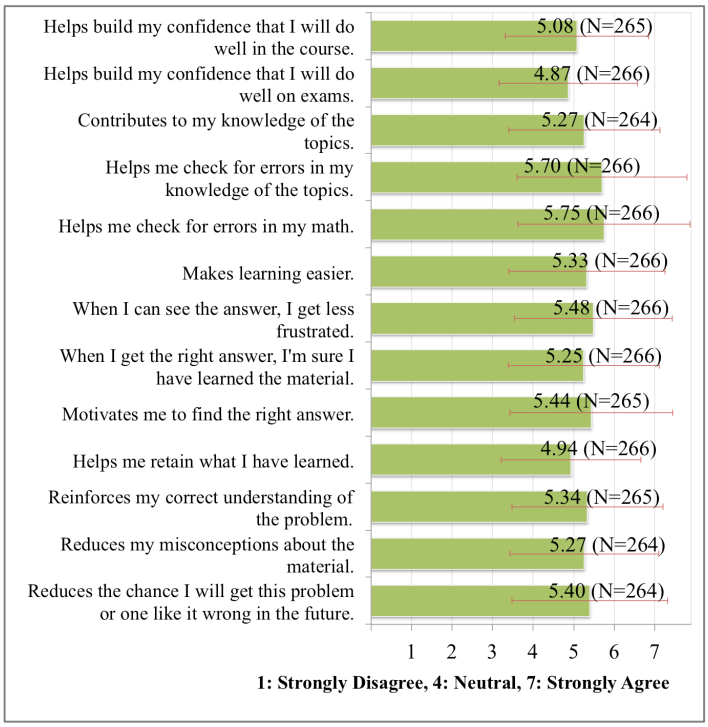


Figure B 3 Scale for students' perception of the utility of the "checkable answers" feature (13 items)

Table B 3 Relationship of students' scale score on "checkable answers" perception and their course performance level

	Mean Scale Score of Perception of Checkable Answer Feature	**F-value**	**P-value**
Students in the high performance group for cumulative grade (N=64)	5.366		
Students in the medium performance group for cumulative grade (N=152)	5.344	0.294	0.746
Students in the low performance group for cumulative grade (N=50)	5.147		
*p <0.05, ** p < 0.01 Performance level: high [90, 100], medium [80, 90), low [0, 80)			

Appendix C Semi-Structured Student Interview Protocol

[Please note: Follow-up, probing questions may build on responses to these general questions]

Identity

1. Are you a freshman? If so, what words best describe what it is like to be a freshman at your university? If not a freshman, what were your previous experiences like as a freshman?
2. What is a day in the life of a freshman in your university like?
3. Overall, what has your semester been like?

Course specific questions

1. Can you tell me about the experience in PHYS101 this semester? How is it going for you?
2. What is a typical PHYS101 class like? A typical homework assignment?
3. How often do you go on the class site on a weekly basis? What areas of the site do you find particularly useful? Not as helpful?
4. How do the online resources compare to traditional resources such as a written textbook, only handing in written problem sets, or labs?
5. Have you ever been in a blended class before? How does PHYS101 compare to blended classes you have previously experienced? How does PHYS101 compare to other face-to-face only classes you have experienced?

a. Definition of "Blended Learning" from the survey: we define blended learning as a class that has a significant online component, but in which the instructor and students still meet face-to-face.

6. Have you been on the discussion board? What is it like? Do you think there is an online community of students in PHYS101?
7. Do you feel as if there is an in-class community?
8. Did you feel that your 3-person team works well together?
9. What were your expectations coming in to the course? Were your expectations met?
10. What is it like to use the materials online, not only the "checkable answers," but all resources, including reading summaries, PDF text book, office hour calendar, Piazza as a discussion forum, grade book, online progress tracker, class slides, in class problems/review problems PDF, etc.? How do you use them?
11. Do you feel PHYS101 has prepared you for future courses in your major?
12. What might you change about the PHYS101 site in the future? [checkable answers feature, videos, reading questions, etc.]

Beliefs about learning/more identity

1. How do you learn best? What have you learned this semester about the way you learn?
2. Did your study strategies change during the semester? How did the online materials influence changes in your study strategies, if they did at all?
3. How do you identify yourself (e.g. female, African American)? What words best describe what it is like to have that identity at your university? Can you tell me about your experiences being a x, y, or z at your university? ?

4. What advice would you consider giving a student from a similar background as you who is enrolling in PHYS101?
    a. Can you tell me why?
    b. Can you tell me more about your particular experiences in PHYS101?
5. How do you describe someone who is smart? Has this definition changed since you came to this university?
6. Do you believe that a person can change how smart he or she is?

Appendix D Problem-Solving Observation Think-aloud Prompt Protocol

We would like you to work the two on-line problems on problem set X. Please do the problems as you normally would. But we would like you to explain to us what you are thinking as you do the problems. We would like to know, for example, why you took the first step you did, why you asked for the answer when you did, and how getting the answer helped you understand the problem. Let us know before you submit each time, we would like to know how confident you are that you got the answer correct.

Paper and pencil will be provided in case you want to use those to solve the problems as well. Let us know if there are other materials you typically use.

Actions or attitudes to make note of:

- Watch for mouse behavior
- Watch for posture and gesturing
- Watch for facial expressions
- Watch for arm crossing/uncrossing
- Watch for checker usage behaviors
- What for techniques used to solve the problem, and whether they match what the problem asks for

Probe or follow-up questions:

- “Can you tell me why you...” <clicking or writing action students have taken but not verbalized>
- “Can you tell me what you’re thinking?” <if students have stopped talking or clicking>

- "You look frustrated <surprised, confused, happy>...can you tell me what you're thinking?" <if students have stopped talking but are expressing non-verbal cues
- "Can you tell me about how often you access this <non-EdX or EdX reference> site? How did you find out about it?"

Important turning points:

- What do students do when they get the question wrong on the first time?
- What do they do when they get the question wrong the 2nd, 3rd, or 4th time?
- What do students do just before submitting an answer?
- Do they express relief, etc. when they submit the answer?
- What do they do when they turn to paper/pencil?
- What do they do when they go back to a previous page?

Before submitting the answer:

- How confident are you that you got the answer correct?

After submit the answer:

- What was the hardest thing about that problem?
- What was the easiest thing?

Appendix E Details for the Behavioral Variables

Table E 1 30 Variables for online homework problems

	Mean	**Median**	**Min**	**Max**
1 Correct Checks	0.92	0.95	0.05	1.77
2 Incorrect Checks	4.69	4.00	0.23	36.73
3 Total Checks	5.60	4.95	0.95	37.32
4 Correct Fraction	0.41	0.39	0.13	0.87
5 First Correct Fraction (before weighting using difficulty levels)	0.19	0.16	0.00	0.69
6 Last Correct Fraction (before weighting using difficulty levels)	0.83	0.88	0.00	0.90
7 Not Attempted	1.16	0	0	17
8 First to Due	69 hours	64 hours	5 hours	9 days
9 Last to Due	58 hours	53 hours	3 hours	8 days
10 First to Last	13 hours	11 hours	4 minutes	61 hours
11 First to Second	3 hours	1 hour	22 seconds	20 hours
12 Interval between Checks	3 hours	2 hours	90 seconds	19 hours
13 Overlap Time	12 hours	8 hours	107 seconds	4 days
14 Weeks Overlap Exists	5.78	6	0	9
15 Activity After Incorrect	0.03	0.02	0.00	0.18
16 Num Session	20.98	20	5	49
17 Time All Sessions	27 hours	26 hours	6 hours	59 hours
18 Avg Session Length	77 minutes	76 minutes	62 minutes	107 minutes
19 Interval between Sessions	4 days	4 days	2 days	13 days
20 Interval within Sessions	4 minutes	4 minutes	21 seconds	13 minutes
21 Video Time	18 minutes	0	0	3 hours
22 Video Num	16.62	0	0	371

Table E 1 continued

	Mean	Median	Min	Max
23 Text Time	49 minutes	29 minutes	0	6 hours
24 Text Num	94.54	45	0	927
25 ClassP Time	18 minutes	6 minutes	0	4 hours
26 ClassP Num	3.62	2	0	62
27 FridayP Time	7 minutes	0	0	2 hours
28 FridayP Num	1.27	0	0	20
29 Exam Time	7 minutes	0	0	3 hours
30 Exam Num	0.84	0	0	26

Table E 2 33 Variables for written homework problems

	Mean	Median	Min	Max
1 Correct Checks	0.79	0.84	0.05	1.65
2 Incorrect Checks	5.07	4.71	0.10	33.14
3 Total Checks	5.85	5.63	0.14	34.22
4 Correct Fraction	0.26	0.25	0.04	0.57
5 First Correct Fraction (before weighting using difficulty levels)	0.000031	0.000027	0.00	0.000153
6 Last Correct Fraction (before weighting using difficulty levels)	0.000292	0.000325	0.00	0.000418
7 Not Attempted	9.57	4	0	57
8 First to Due	52 hours	49 hours	-23 hours (started after the due time)	7 days
9 Last to Due	44 hours	41 hours	-23 hours (started after the due time)	6 days
10 First to Last	13 hours	11 hours	4 minutes	61 hours
11 First to Second	9 hours	7 hour	7 minutes	30 hours

Table E 2 continued

	Mean	Median	Min	Max
12 Interval between Checks	102 minutes	84 minutes	2 minutes	17 hours
13 Activity After Incorrect	0.02	0.02	0.00	0.16
14 Num Session	39.90	40	3	77
15 Time All Sessions	69 hours	70 hours	5 hours	5 days
16 Avg Session Length	104 minutes	103 minutes	67 minutes	161 minutes
17 Interval between Sessions	59 hours	51 hours	25 hours	28 days
18 Interval within Sessions	6 minutes	6 minutes	43 seconds	24 minutes
19 Video Time	22 minutes	1 minute	0	4 hours
20 Video Num	20.25	2	0	1589
21 Text Time	140 minutes	93 minutes	0	14 hours
22 Text Num	252.90	139	0	1953
23 ClassP Time	42 minutes	25 minutes	0	6 hours
24 ClassP Num	8.02	4	0	80
25 FridayP Time	16 minutes	33 seconds	0	3 hours
26 FridayP Num	2.76	1	0	39
27 Exam Time	11 minutes	0	0	1 hours
28 Exam Num	1.29	0	0	16
29 Correct Fraction before Correct Steps	0.87	0.90	0.23	1.00
30 Incorrect Fraction before Correct Steps	0.04	0.03	0.00	1.00
31 Skipped Fraction before Correct Steps	0.09	0.06	0.00	0.68
32 Problems Containing Incorrect Steps	0.07	0.06	0.00	0.60
33 Problems Containing Skipped Steps	0.13	0.10	0.00	0.75

Appendix F Pair-wise Correlation Matrices of Behavioral Variables

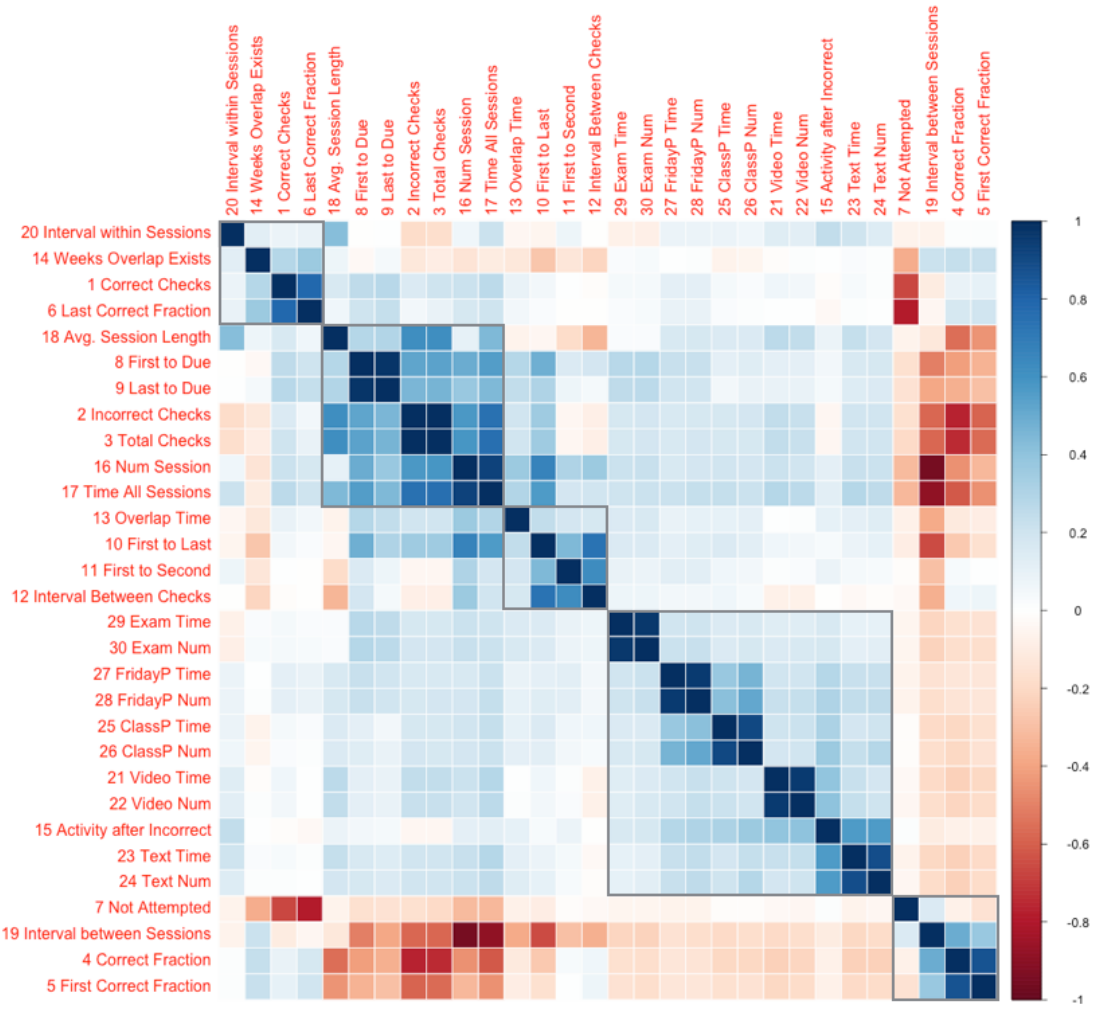


Figure F 1 Correlation matrix of all 30 behavioral variables for online homework problems

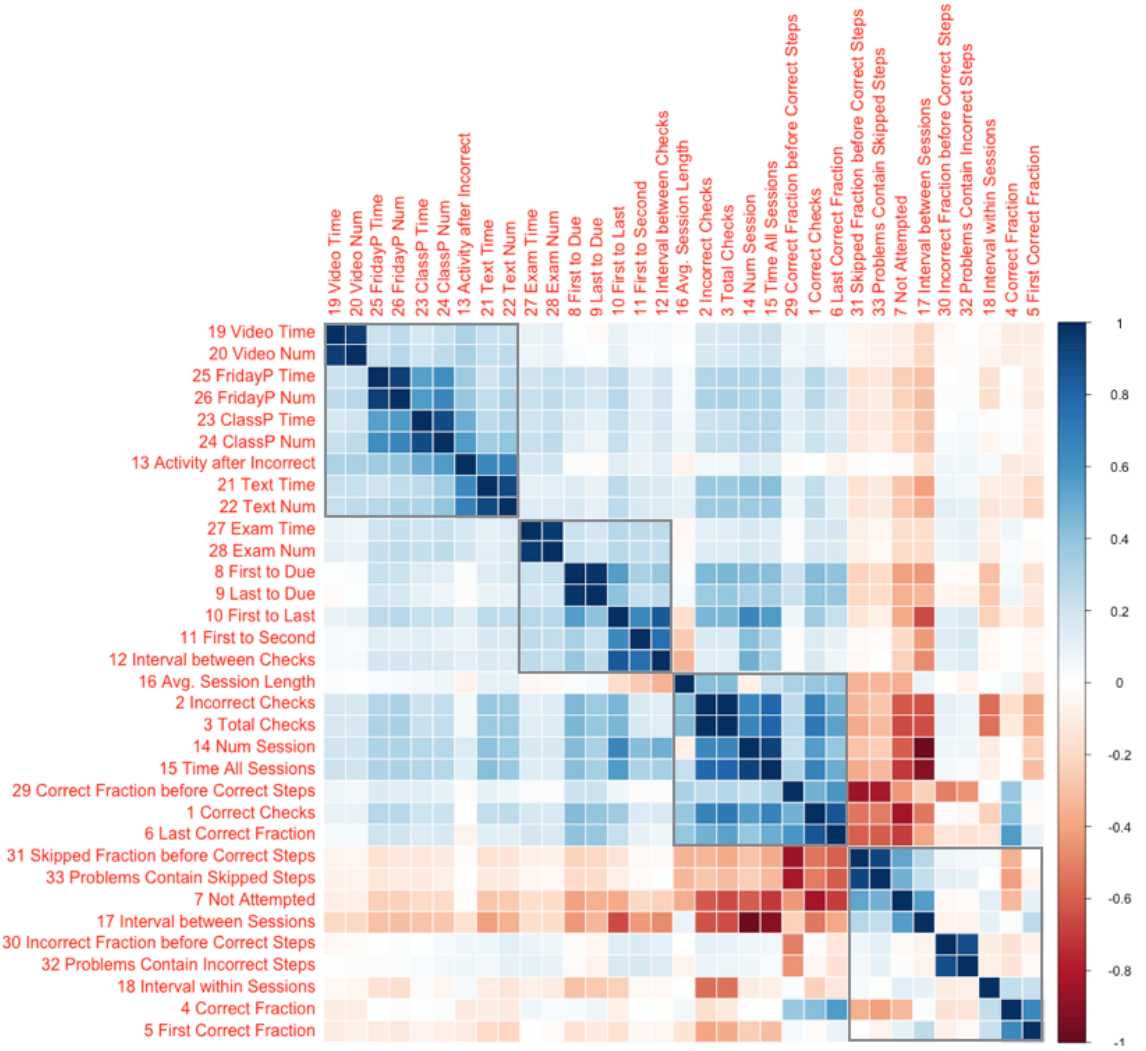


Figure F 2 Correlation matrix of all 33 behavioral variables for written homework problems

Appendix G Centroids of Student Clusters with All Variables

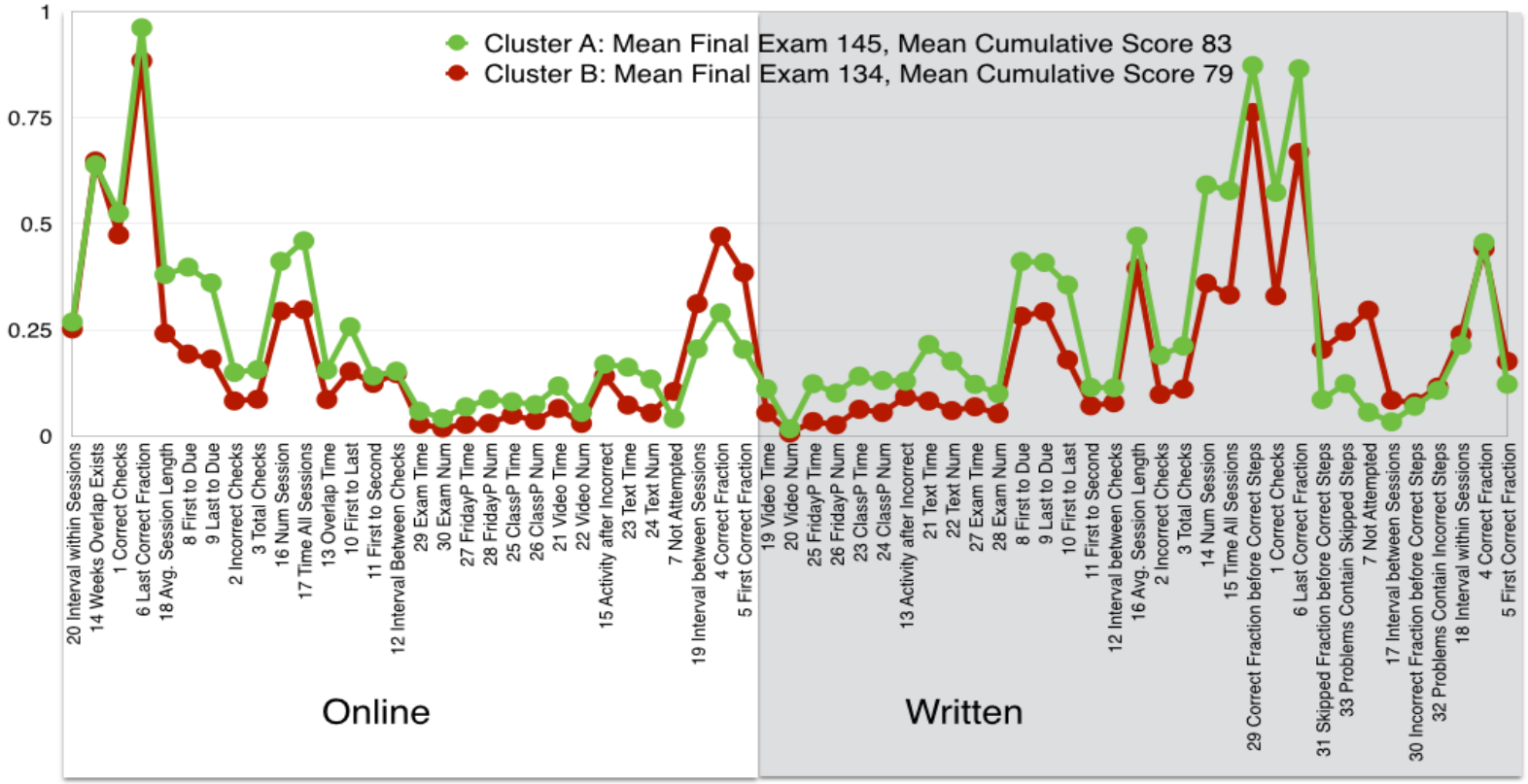


Figure G 1 Centroids of two student clusters with all variables for online and written homework (cluster A: 316 students; cluster B: 157 students)

VITA

VITA

Xin Chen
School of Engineering Education, College of Engineering, Purdue University

Education
Ph.D., Engineering Education, Purdue University, West Lafayette, Indiana, 2010-2015
B.S., Electrical Engineering, East China Normal University, Shanghai, China, 2006-2010
Exchange Student, Electrical Engineering, Shanghai Jiaotong University, Shanghai, China, 2007-2008

Research Interests
Educational data mining, student behaviors in online and blended learning environment, social media data analytics, user experience research

Data Science Experience
Data Science Fellow, Insight Data Science, Palo Alto, California, 2015
Data Science Fellow, Eric & Wendy Schmidt "Data Science for Social Good" Summer Fellowship, The University of Chicago, Chicago, Illinois, 2014

Honors and Awards
Outstanding Research Award, School of Engineering Education, Purdue University, 2014
Best Poster Award, International Conference on Learning Analytics and Knowledge (LAK), 2014